Latent class models for density estimation

van der Palm, D.W.

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Latent Class Models for Density Estimation

Daniël W. van der Palm

Daniël Willem van der Palm

Latent Class Models for Density Estimation

Thesis, Tilburg University
October 31, 2013

ISBN: 978-90-5335-772-9

Cover: D.W. van der Palm with help of Robert Kanters
Print: Ridderprint BV

Latent Class Models for Density Estimation, with Applications in Missing Data Imputation and Test-Score Reliability Estimation

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University, op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit

op vrijdag 6 december 2013 om 14.15 uur

door

Daniël Willem van der Palm,

geboren op 19 maart 1985 te Capelle aan den IJssel

Promotores: Prof. dr. J. K. Vermunt

Prof. dr. K. Sijtsma

Copromotor: Dr. L. A. van der Ark

Overige leden van de Promotiecommissie:

Prof. dr. J. A. P. Hagenaars

Prof. dr. S. van Buuren

Prof. dr. H. L. J. van der Maas

Dr. W. H. M. Emons

Dr. T. de Waal

Contents

Chapter 1

Introduction

Latent class analysis (Lazarsfeld, 1950) is frequently used in the social and behavioral sciences as a method to identify meaningful subgroups in multivariate categorical data, and is commonly referred to as a categorical data analogue to factor analysis (McCutcheon, 1987). Many substantive concepts in the social sciences cannot be directly measured. Examples are personality subtypes (e.g., introverted, extraverted) and diagnostic subcategories in psychiatric patients (e.g., healthy, depressed, schizophrenic). Therefore, observable indicator variables must be used as an indirect measure of the concept one wishes to measure. For example, the statement "I see myself as talkative" could be an indicator variable for the concept of extraversion, with five possible ordered answers ranging from "strongly disagree" to "strongly agree". In latent class analysis, it is postulated that a nominal latent variable underlies the responses to the indicator variables, and that the associations between the indicator variables can be explained by their dependence on the latent variable. After having estimated a latent class model, a substantive interpretation can be given to the latent variable by examining the associations of each latent class with the indicator variables. However, in this dissertation, the latent class model is used as a density estimation tool, which differs from using the latent class model as a substantive model. Because the latent class model is primarily known as a substantive model, and the use of the latent class model as a density estimation tool is a more recent development, we discuss both ways of using the latent class model. In this way, we can also clarify why using the latent class model as a density estimation tool is more straightforward than using the model for substantive analysis, and why several technical aspects of the latent class model need to be considered in the latter type of use, but not in the former.

The latent class model is a mixture model and, similar to mixture models in general, the association structure of a set of variables, which is the density, can be modeled using a finite mixture of simpler densities (McLachlan & Peel, 2000). More specifically, each latent class has a class-specific multinomial density that describes the probabilities of giving a specific response to each of the indicator variables; these are the conditional response probabilities. Furthermore, within each latent class the responses to the indicators are assumed

to be statistically independent from one another, also known as the local independence assumption (Lazarsfeld, 1950). Latent class analysis is similar to cluster analysis, which is defined as the classification of similar objects into groups, without prior knowledge of the number of groups and what form they have (Kaufman & Rousseeuw, 1990). However, in contrast to standard clustering techniques, latent class clustering is model-based. That is, a specific statistical model defined as the mixture of multinomial densities is used to describe the associations between the latent variable and the indicator variables, and to cluster respondents. Maximum likelihood estimates for the parameters of the latent class model are usually obtained by means of the expectation maximization algorithm (Goodman, 1974; Dempster, Laird, & Rubin, 1977), and/or a Newton Raphson algorithm.

In general, density estimation can be described as a statistical procedure with the goal to approximate the distribution of a set of variables at the population level, based on a sample drawn from this population. Thus, the idea is that the underlying distribution can be inferred from the associations between the variables in the sample. Once the theoretical density of a dataset has been estimated, it can be used for various applications. For example, density estimation using a latent class model has been used for multiple imputation of missing categorical data (Gebregziabher and DeSantis, 2010; Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008), to smooth large sparse contingency tables (Linzer, 2011), and to estimate test-score reliability (Van der Ark, Van der Palm, & Sijtsma, 2011). Before we discuss the specific differences between using the latent class model as a density estimation tool and as a substantive model, we consider the practical problem that led to the development of the latent class model as a density estimation tool.

A well known and established model for categorical data is the log-linear model (Bishop, Fienberg, & Holland, 1975; Agresti, 1990; Hagenaars, 1990). The log-linear model can also be used as a density estimation tool. However, due to computational issues a log-linear model can only be estimated for datasets with a small number of variables; the number of cells in the contingency table that has to be evaluated quickly becomes too large. Furthermore, it is common practice to use a saturated log-linear model for density estimation because this model is able to capture all possible associations. For a saturated log-linear model, computational issues are encountered even sooner because a saturated model implies that all response patterns that are theoretically possible have to be stored and evaluated, and that number may be too large to store in computer memory. For example, the number of cells to evaluate exceeds one million for 20 dichotomous variables ($2^{20} = 1{,}048{,}576$), and one billion for 30 dichotomous variables, 19 trichotomous variables, or 13 variables with five

categories. The computational problems of the log-linear model limit its practical usefulness and a more practical alternative was needed, which gave rise to research on the latent class model as a density estimation tool.

Compared to the log-linear model, the main advantage of the latent class model as a density estimation tool is that the latter model can handle a very large number of variables. There are two main characteristics of the latent class model that allow the inclusion of a very large number of variables without impairing the computational capacity of the model. First, due to the local independence assumption the latent class model has a relatively simple structure, which is a finite mixture (i.e., weighted average) of independent multinomial densities. Second, the latent class model does not require an a priori specification of which associations the model should take into account. If a sufficient number of latent classes is chosen, the latent class model is able to capture first, second, and higher order associations, and typically a sufficient number of latent classes does not constitute a saturated model. Hence, using a single model specification (i.e., a certain number of latent classes), a latent class model is able to capture different association structures in different datasets and, therefore, requires relatively few parameters. In contrast, for the log-linear model one must exactly specify a priori which associations the model should include. Hence, the latent class model is highly flexible.

If the latent class model is used for substantive purposes, there are several criteria and potential problems that must be considered. The number of latent classes one chooses should typically follow from substantive theory and the number of latent classes is preferred to be small to facilitate interpretation. Furthermore, latent class models may be unidentified if the number of latent classes is large relative to the number of observed variables (Goodman, 1974). An unidentified model means that there is no unique set of parameters associated with the global maximum of the data log-likelihood and, consequently, a meaningful substantive interpretation of the model is difficult if not impossible. Lastly, it is important to find the set of parameter estimates for which the likelihood of the data is at a global maximum; the probability of finding this specific set of parameter estimates decreases as a function of the specified number of latent classes. Thus, it is possible that the estimation procedure of a latent class model yields a set of model parameter estimates associated with a local maximum of the data. However, if the latent class model is used as a density estimation tool, the number of latent classes does not follow from substantive theory, and interpretation of the parameter estimates does not make sense. Therefore, a large number of latent classes is not problematic and the latent class model does not have to be identified. Furthermore, ending up in a local

maximum is not problematic as the local solution is typically nearly as good as the global maximum solution (Vermunt et al., 2008). The single criterion for a latent class model as a density estimation tool is whether the model captures all relevant associations amongst the variables. Yet, the question remains which associations are relevant. We attempt to answer this question in this dissertation.

Because the latent class model can handle a large number of variables, applications that make use of latent-class based density estimation are also widely applicable. The wide applicability was one of the main reasons why Vermunt et al. (2008) investigated the performance of the latent class model as an incomplete-data method.

The Latent Class Model as an Incomplete-Data Method

Incomplete-data is a frequently encountered phenomenon in the social sciences that researchers must deal with before the statistical analysis of interest can be performed. Multiple imputation (MI; Rubin, 1987) has become widely recognized as a sound approach to address incomplete-data problems. The four steps of MI can be outlined as follows: (1) define and estimate an imputation model to capture the associations in a dataset, (2) use the imputation model to obtain *m* predicted values for substitution of each missing value, creating *m* completed datasets, (3) estimate the statistical model of interest on each of the *m* datasets, obtaining *m* sets of estimated model parameters, and (4) pool the results of the *m* analyses to obtain the final results.

The basic idea of MI is to fill in every missing value using an imputation value that is consistent with the association structure of a sample dataset (Rubin, 1987). Thus, if the imputation model has a sufficient fit to the data then the imputation values that are obtained using the imputation model will not distort the association structure present in the data (i.e., behave neutrally). Furthermore, the variation in the imputation values across the *m* completed datasets should correctly reflect two sources of uncertainty: The imputed values are uncertain because they are actually missing values and the estimated model parameters are uncertain because of sampling error. The final goal of MI is to allow researchers to estimate the substantive model in such a way that the missing data do not affect the results in an undesirable way. More specifically, there are two practical criteria that must be considered when evaluating an MI method: (1) the estimated parameters of the substantive model should not be distorted by imputation (i.e., a bias criterion), and (2) the standard errors of the

parameters of the substantive model should correctly reflect the uncertainty due to the missing values (i.e., a bias of standard errors criterion).

Vermunt et al. (2008) introduced MI using a latent class model (MILC). In a simulation study the authors compared four incomplete-data methods: Two variants of MILC, maximum likelihood for incomplete data (MLID; Dempster, Laird, & Rubin, 1977; Little & Rubin, 2002), and MI using a log-linear model (MILL; Schafer, 1997). The first variant of MILC used AIC3 to choose the number of latent classes, and the second variant of MILC used a fixed large number of latent classes. Vermunt et al. (2008) found that MILC using a fixed large number of latent classes produced the least bias in parameter estimates and standard errors of the substantive model, and had a performance comparable to MLID and MILL. However, MILC had not yet been compared to other methods such as multivariate imputation using chained equations (MICE; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; Van Buuren & Oudshoorn, 2011). Furthermore, in the study by Vermunt et al. (2008) the influence of sampling error, sample size, and complexity of associations on the performance of the latent class model as a multiple imputation method had not been taken into account. Chapter 2 addresses these two issues.

The Divisive Latent Class Model

Although density estimation using a latent class model is widely applicable, a practical issue remains. Before the density estimate can be obtained, one must first determine how many latent classes are sufficient to obtain a precise density estimate, capturing all associations. A typical model fit strategy to find the number of latent classes that is sufficient is to fit a latent class model with one class, two classes, and so on, until the optimal latent class model is found. This process may require an excessive computation time, especially for datasets with a large number of variables and respondents. Furthermore, if the researcher has to estimate and compare a large number of latent class models, the amount of work may be an obstacle to continue and, moreover, such a procedure is relatively prone to human error. Therefore, the question arises whether there might be a more efficient way to estimate a latent class model. In Chapter 3, the divisive latent class (DLC) model is introduced that addresses the problems of excessive computation time, laboriousness, and error proneness.

The DLC model estimation procedure constitutes a top-down clustering of respondents into latent classes. It is obtained by estimating a series of one-class and two-class models. Thus, the best fitting latent class model is produced in a single run and, in contrast to a

standard latent class model, each step during the estimation procedure builds on results of the previous steps. For this reason, the estimation of a DLC model is much faster compared to the estimation of standard latent class models. In addition to addressing the practical issues of the standard latent class model, we investigated which model fit criteria should be used for the DLC model in the context of density estimation.

The Divisive Latent Class Model as an Incomplete-Data Method

As mentioned before, using a latent class model with a large number of latent classes is not problematic in the context of density estimation. Therefore, in the third chapter, we investigated whether model parsimoniousness could be traded for computational efficiency and introduced the DLC model as a density estimation tool. Because a DLC model is much faster than the standard latent class model, the question also arises how well the model would perform as an incomplete-data method.

In the fourth chapter, the DLC model is applied to the problem of missing data. Because of technological developments such as the Internet, the amount of available data is becoming ever greater (e.g., data archives of institutions such as GESIS, and data collectible from social media such as Twitter). Therefore, more and more datasets contain a very large number of respondents and variables. Using a standard latent class model for such datasets may require excessive computation time. As for density estimation in general, the DLC model may alleviate the practical burden of excessive computation time it places on software and researchers. However, it has not yet been investigated which model fit strategy one should use for the DLC model as an incomplete-data method. Chapter 4 addresses this question using three simulation studies; two with an artificially defined population model, and one with a population model based on the observed associations in a real dataset.

The Divisive Latent Class Model as a Test-Score Reliability Estimation Method

Another important topic in the social and behavioral sciences is test-score reliability estimation. If highly important decisions are based on test scores, it is important that test scores are reliable. In the classical test theory sense, reliability concerns the question of whether test scores are repeatable to a high degree across hypothetical, independent replications of the same test (Lord & Novick, 1968, p. 61). The product-moment correlation between two independent administrations of the same test to the same sample of persons

provides an estimate of test-score reliability. However, in real life one usually tests the same persons only once and, therefore, a reliability estimate based on repeated testing is practically impossible. For this reason, many methods have been proposed to estimate test-score reliability on the basis of the data obtained in one test administration (Cronbach, 1951; Guttman, 1945; Lord & Novick, 1968).

Researchers and practitioners commonly resort to using classical lower bounds to the reliability such as coefficient alpha (Cronbach, 1951), and coefficient lambda2 (Guttman, 1945). It is well known that alpha and lambda2 require rather restrictive assumptions to be equal to the reliability. The data must be unidimensional, and the test items must be essentially tau equivalent; in practice these assumptions never completely hold. Van der Ark et al. (2011) introduced the latent class reliability coefficient (LCRC) which does not require such restrictive assumptions, and showed that LCRC yields practically unbiased reliability estimates. The study by Van der Ark et al. (2011) contained one condition concerning multidimensional data for which LCRC performed very well in terms of bias relative to the population reliability. However, it was not yet known whether this result would generalize to a wider range of scenarios involving multidimensional data. In addition, it was not yet known how the divisive latent class model would perform as an adaptation of LCRC. In Chapter 5, the combination of LCRC and the divisive latent class model is investigated as a reliability estimation method for multidimensional educational test data.

The chapters were all written as separate articles intended for publication in academic journals. The contents of each chapter were kept as close to the original articles as possible. Therefore, there may be some overlap between the chapters, and notation may change across chapters.

Chapter 2

A Comparison of Incomplete-Data Methods for Categorical Data

Abstract

We studied four methods for handling incomplete categorical data in statistical modeling: (1) maximum likelihood estimation of the statistical model with incomplete data, (2) multiple imputation using a log-linear model, (3) multiple imputation using a latent class model, (4) and multivariate imputation by chained equations. Each method has advantages and disadvantages, and it is unknown which method should be recommended to practitioners. We reviewed the merits of each method and investigated their effect on the bias and stability of parameter estimates and bias of the standard errors. We found that multiple imputation using a latent class model with many latent classes was the most promising method for handling incomplete categorical data, especially when the number of variables used in the imputation model is large.

This chapter has been accepted for publication as: Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2013). A comparison of incomplete-data methods for categorical data. *Statistical Methods in Medical Research*, in press.

2.1 Introduction

This chapter discusses methods to handle incomplete categorical data. Many medical studies deal solely with analyzing categorical data and, consequently, the statistical model that is used to analyze the data (from here on referred to as the *substantive model*) is also tailored to categorical data. For example, predictors of reduced length of hospital stay were studied using logistic regression (Kurian, Gallagher, Cheeyandira, & Josloff, 2010), determinants of caregivers' health were studied using log-linear modeling (Zhu, Walter, Rosenbaum et al., 2006), and the effectiveness of the World Health Organization Disability Assessment Schedule II was investigated using a nonparametric item response analysis (Luciano, Ayuso-Mateos, Aguado et al., 2010). A frequently encountered problem is that the data are incomplete, which prevents a straightforward statistical analysis; a researcher should handle this problem appropriately. Klebanoff and Cole (2008) found that the majority of applied researchers resort to ad-hoc methods such as complete-case analysis or pair-wise deletion, which may lead to biased statistical results and reduced power (Little & Rubin, 2002; Schafer, 1997). For handling incomplete continuous data, adequate alternatives have been proposed, extensively researched (Schafer & Graham, 2002), and implemented in major software packages such as SPSS (SPSS Inc., 2011) and SAS (SAS Inc., 2011). Hence, there is no need for applied researchers to resort to ad-hoc methods in case of continuous data.

Incomplete data methods for categorical data have not yet been crystallized out, and it is unknown which method should be recommended to practitioners. Ideally, an incomplete-data method should meet three criteria. For the substantive model, it should produce parameter estimates (i) that are unbiased, (ii) that are stable in order to avoid unnecessary loss of power in the statistical analysis, and (iii) that have standard errors correctly reflecting the uncertainty due to missing data. Ideally, these criteria should be met for datasets with both small and large numbers of variables, sample sizes, and percentages of incomplete data, and for both simple and complex associations in the data.

With respect to these criteria, two incomplete-data methods for categorical data are especially promising: Multiple imputation *using latent class analysis* (MILC; Gebregziabher & DeSantis, 2010; Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008) and *multivariate imputation using chained equations* (MICE; Van Buuren, 2007; Van Buuren, Brand, Groothuis-Oudshoorn et al., 2006; Van Buuren & Groothuis-Oudshoorn, 2011). Both methods also have the practical advantage that they can easily handle datasets containing a large number of variables and respondents. However, researchers having incomplete

categorical data cannot yet readily apply MILC and MICE because there are various unresolved issues (explained hereunder). The impact of these issues on the three criteria for substantive models is unknown. In this study, we discuss two reasonable options for the unresolved issues for both MILC and MICE, and investigate to which degree they meet the three criteria, so as to decide which incomplete-data method should be selected for categorical data. Multiple imputation *using a log-linear model* (MILL; Schafer, 1997) and *maximum likelihood for incomplete data* (MLID; Allison, 2001; Arbuckle, 1997; Dempster, Laird, & Rubin, 1977; Little & Rubin, 2002; also known as full information maximum likelihood) are used as benchmarks. MILL is known to produce unbiased parameter estimates (Ezzati-Rice, Johnson, Khare et al., 1995; Schafer, 1997; Schafer, Ezzati-Rice, Johnson et al., 1998) but can only handle a small number of variables; MLID is known to be asymptotically unbiased, but may run into difficulties as the number of variables becomes very large (Vermunt et al., 2008).

The remainder of this chapter is organized as follows. First, we briefly discuss the four incomplete-data methods. For both MILC and MICE we discuss two variants, resulting in six incomplete-data methods in total. Second, we compare the advantages and disadvantages of the methods in a theoretical discussion. Third, we present the results of two simulation studies. In Study 1, for dichotomous data, we compared MILC, MICE, MILL, and MLID with respect to the three criteria. In Study 2, for trichotomous data, we compared MILC, MICE, and complete-case analysis with respect to the three criteria. Fourth, we applied MLID, MILC, MICE, and complete-case analysis to a medical dataset. Finally, we give recommendations based on the theoretical discussion and the two simulation studies.

2.2 Incomplete-Data Methods

2.2.1 Incomplete data

Let $\mathbf{Y} = (Y_1, \ldots, Y_J)$ denote the scores on the J variables, and let $\boldsymbol{\theta}$ be the generic notation for the vector of unknown parameters of the joint distribution of $\mathbf{Y}$, denoted $P(\mathbf{Y}; \boldsymbol{\theta})$. To distinguish specific models Greek letters other than $\boldsymbol{\theta}$ may also be used to denote parameter vectors. Note that Y_j may be either a predictor variable or an outcome variable depending on the substantive model. If confusion arises, we add the superscripts p and o to indicate that a variable serves as a predictor variable or outcome variable, respectively. $\mathbf{Y}$ may contain missing values, and the objective is to deal with them appropriately.

Most incomplete-data methods, including the ones considered in this chapter, assume that the mechanism that caused the missing values is *ignorable* (Allison, 2001), which means that two conditions should hold. First, the parameters that govern the missing data process must be unrelated to the parameters to be estimated, which is a rather unrestrictive assumption (Little & Rubin, 2002). Second, the data must be *missing at random* (MAR), which means that whether or not a score is missing only depends on scores observed in the study. If, after conditioning on all observed data, the missingness depends on missing values of variables included in the study or on variables not included in the study, MAR is violated and, as a result, the missingness mechanism is non-ignorable. Non-ignorable missingness may cause biased parameters in the substantive model (first criterion). Apart from special studies with planned missingness (Schafer, 1997), MAR is unlikely to hold in practice, and it is impossible to test whether the MAR assumption holds for a particular dataset (Schafer & Graham, 2002). Therefore, the degree to which MAR is violated (i.e. the degree to which the observed scores cannot explain the missingness mechanism) becomes important: If the violation of MAR becomes more severe, the parameter bias in the substantive model is likely to increase. If the number of variables in a dataset increases, the degree to which the variables can explain the missingness mechanism is also likely to increase. Hence, if an incomplete-data method can handle a large number of variables, and if a large number of variables is available, the violation of MAR will most likely be less severe and the missingness mechanism is more likely to be ignorable. This notion (Schafer, 1997) plays an important role in our evaluation of incomplete-data methods, and will be referred to as *Schafer's notion on the number of variables*.

2.2.2 Description of Incomplete-Data Methods

2.2.2.1 *Maximum likelihood for incomplete data.*

MLID is a well known and documented method to obtain parameter estimates and standard errors in the presence of missing data (Little & Rubin, 2002). MLID constitutes estimating the parameters of the substantive model and their standard errors, using all observed data. For example, when studying predictors of reduced length of hospital stay using logistic regression (Kurian et al., 2010), MLID can be used to estimate the logistic regression model using all observed data. No further action is required; the obtained parameter estimates and standard errors can be directly interpreted. The substantive model can be an asymmetric model such as a logistic regression model or an item response theory model, which describe the conditional distribution of the outcome variables given the predictor variables $P(\mathbf{Y}^o|\mathbf{Y}^p;\boldsymbol{\theta})$, or a

symmetric model, such as a log-linear model, latent class model, or canonical correlation model, which describe the joint distribution of all variables $P(\mathbf{Y}; \boldsymbol{\theta})$. MLID assumes that the missingness mechanism is ignorable. For categorical data, specialized software is usually required to conduct MLID, such as LEM (Vermunt, 1997) or Mplus (Muthén & Muthén, 2010).

2.2.2.2 *Multiple Imputation*

Multiple imputation consists of creating *m* completed datasets by replacing the missing values in the data with plausible values *m* times. These plausible values replacing the missing values are called the imputed values. The statistical model that generates imputed values is referred to as the *imputation model*. After the multiple imputation, on each of the *m* completed datasets a substantive model is estimated, and the *m* sets of parameter estimates and standard errors are combined into a single set. Most researchers use $m = 5$, but this value is currently debated (White, Royston, & Wood, 2010). Using multiple imputation allows for separating the missing data handling and the substantive analysis; a researcher can estimate substantive models as if there had been no missing data, or distribute the completed data to other researchers for further analysis.

Multiple imputation starts in the same way as MLID for symmetric models: A statistical model is estimated describing the joint distribution $P(\mathbf{Y}; \boldsymbol{\theta})$. Rather than a substantive model, this model is an imputation model for obtaining imputed values from $P(\mathbf{Y}; \boldsymbol{\theta})$. For example, when studying predictors of reduced length of hospital stay using logistic regression (Kurian et al., 2010), a log-linear model describing the joint distribution of both predictor variables and reduced length of hospital stay may be used as an imputation model to generate imputed values replacing the missing data *m* times. After the multiple imputation, logistic regression analysis can be conducted on the completed datasets.

One must account for the fact that the imputed values are not observed and, therefore, uncertain. There are two sources of uncertainty (Rubin, 1987). Firstly, the estimated parameters of the imputation model are uncertain; this uncertainty is expressed by their standard errors. Secondly, there is uncertainty due to sampling variability when drawing imputed values from $P(\mathbf{Y}; \boldsymbol{\theta})$. To account for parameter uncertainty, for each of the *m* datasets, a different set of parameters of the imputation model is used. In a Bayesian framework, the *m* sets of parameters of the imputation models are random draws from $P(\boldsymbol{\theta}|\mathbf{Y})$, the distribution of the parameters given the data (Little & Rubin, 2002). In a frequentist framework, the *m* sets of parameters are estimated using *m* nonparametric

bootstrap samples of the data (Vermunt et al., 2008). A nonparametric bootstrap sample consists of randomly drawing a new sample of *N* observations with replacement (Efron & Tibshirani, 1993). To reflect uncertainty due to sampling variability, the replacement of missing values is done *m* times, yielding *m* completed datasets. The three multiple imputation methods for categorical data discussed in this paper differ in the way that they describe the joint distribution, $P(\mathbf{Y};\ \boldsymbol{\theta})$, and how they account for parameter uncertainty. MILL is discussed briefly because this method is ready for use; MILC and MICE are described in more detail so as to allow the discussion of the specific options these methods offer.

Multiple imputation using a log-linear model. MILL uses a log-linear model as the imputation model. Let the parameters of the log-linear model be denoted $\boldsymbol{\lambda}$; the saturated log-linear model for dichotomous responses can be written as

$$\log P(\mathbf{Y};\boldsymbol{\lambda}) = \lambda + \sum_{i=1}^{J}\lambda_i Y_i + \sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\lambda_{ij}Y_iY_j + \cdots + \lambda_{1,2,\ldots,J}Y_1Y_2\ldots Y_J \qquad (2.1)$$

The joint distribution is obtained by taking the exponential of the right-hand side of eq. (2.1). Typically, a saturated log-linear model is used to obtain imputation values because it captures all possible associations in the data; therefore, it is the gold standard for multiple imputation of categorical data (Vermunt et al., 2008). If higher-order interaction terms are omitted, the approximation of the joint distribution by the log-linear model may deteriorate. MILL can, for example, be conducted using software packages CAT (Schafer, 1997) or Latent GOLD 4.5 (Vermunt & Magidson, 2008), which utilize a Bayesian and a nonparametric bootstrap approach, respectively, to account for parameter uncertainty.

Multiple imputation using a latent class model. MILC uses a latent class model to estimate the joint distribution of the variables in the data. Let *X* denote a discrete latent variable with K latent classes, indexed by k $(k = 1,\ldots,K)$. Let $\boldsymbol{\pi}$ denote the vector of parameters of the latent class model; $\boldsymbol{\pi}$ can be divided into $\boldsymbol{\pi}_x$, the latent class proportions, and $\boldsymbol{\pi}_y$ the conditional response probabilities. Under a latent class model, joint distribution $P(\mathbf{Y};\boldsymbol{\pi})$, has the following form (Goodman, 1974; Lazarsfeld, 1950; Vermunt & Magidson, 2004):

$$\begin{aligned} P(\mathbf{Y};\boldsymbol{\pi}) &= \sum_{k=1}^{K} P(X=k;\boldsymbol{\pi}_x)P\big(\mathbf{Y}|X=k;\boldsymbol{\pi}_y\big) \\ &= \sum_{k=1}^{K} P(X=k;\boldsymbol{\pi}_x)\prod_{j=1}^{J} P\big(Y_j|X=k;\boldsymbol{\pi}_y\big). \end{aligned}$$

If the number of latent classes is sufficiently large, a latent class model correctly picks up the first, second, and higher order moments of the response variables, as is the case with all forms of mixture models (McLachlan & Peel, 2000). It is unknown how many latent classes are sufficient for a good approximation of the joint distribution. Vermunt et al. (2008) argued that it is better to have too many than too few latent classes. Therefore, out of three selection criteria, Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and AIC3 (Bozdogan, 1993), they suggested using AIC to select the number of latent classes because it yields the largest number of latent classes. Hence, letting AIC determine the number of latent classes, abbreviated *MILC (AIC)*, is the first option for MILC. However, it is expected that an even larger number of latent classes can further improve the approximation of the joint distribution. Having a relatively large number of latent classes, abbreviated *MILC (Large)*, is the second option for MILC. MILC can be applied using Latent GOLD, which uses the nonparametric bootstrap to account for parameter uncertainty.

Multivariate imputation using chained equations. MICE is a fully conditional specification method (Van Buuren & Groothuis-Oudshoorn, 2011), which specifies the imputation model on a variable-by-variable basis using a separate conditional distribution for each incomplete variable. Let $\mathbf{Y_{-j}}$ denote the scores on all variables except Y_j. MICE reduces the problem of finding one J-dimensional joint distribution $P(\mathbf{Y};\boldsymbol{\theta})$ to finding J univariate conditional distributions $P(Y_1|\mathbf{Y}_{-1};\boldsymbol{\theta}), \ldots, P(Y_J|\mathbf{Y}_{-J};\boldsymbol{\theta})$ (Van Buuren, 2007; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; Van Buuren & Groothuis-Oudshoorn, 2011). Conditional distribution $P(Y_j|\mathbf{Y}_{-j};\boldsymbol{\theta})$ is used for imputation of $Y_j (j = 1, \ldots, J)$. Under certain conditions, a draw from each of the J conditional distributions is equivalent to a single draw from the joint distribution (Van Buuren, 2007), but it is not guaranteed. Results from simulation studies (Drechsler & Rassler, 2008; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) suggest that the problem is unlikely to be serious in practice.

MICE starts with replacing missing values of the variables by draws from their respective marginal distributions. Next, in an iterative process, the imputed values are updated variable by variable using the univariate conditional distributions. When Y_j is imputed, the other variables act as predictor. If the joint distribution that is defined by the J conditional distributions exists then this iterative process is a Gibbs sampler (Van Buuren, 2007) and converges to the joint distribution of the J variables. Often, as little as 10 to 20 iterations are required.

The imputation model describing the conditional probabilities $P(Y_1|\mathbf{Y}_{-1};\boldsymbol{\theta}), \ldots, P(Y_J|\mathbf{Y}_{-\mathrm{J}};\boldsymbol{\theta})$ can be any appropriate regression model depending on the nature of the outcome variable (Agresti, 1990): Linear regression in combination with predictive mean matching, logistic regression, polytomous regression, and nonlinear regression. We focused on two imputation models; the first one being logistic regression (abbreviated MICE (LOG)) which is the default method in the R-package MICE (Van Buuren & Groothuis-Oudshoorn, 2011) for dichotomous outcome variables (for Study 2, polytomous regression is used, which is the extension of MICE (LOG) to variables with more than 2 categories; for details see, e.g. Van Buuren et al., 2011). Let $\boldsymbol{\beta}$ denote the vector of parameters for the logistic regression model. MICE (LOG) models conditional distribution $P(Y_j|\mathbf{Y}_{-\mathrm{j}};\boldsymbol{\beta})$ as

$$\mathrm{logit}\left[P(Y_j|\mathbf{Y}_{-j};\boldsymbol{\beta})\right] = \beta_0 + \beta_1 Y_1 + \cdots + \beta_{j-1} Y_{j-1} + \beta_{j+1} Y_{j+1} + \cdots + \beta_J Y_J.$$

We also considered linear regression in combination with predictive mean matching (abbreviated MICE (PMM)). The first step of MICE (PMM) is to obtain a predicted value by means of linear regression in which all other variables serve as predictors. In the second step, the respondent that has the most similar predicted value as well as an observed value on the variable that is being imputed is selected as the nearest neighbor. Subsequently, the observed value of this nearest neighbor is used as the imputation value for the respondent with a missing value.

Parameter uncertainty is accounted for in a Bayesian framework; a new set of parameters is drawn from its posterior distribution for the construction of each imputed dataset. More specifically, the MICE algorithm involves iteratively sampling parameter values $\boldsymbol{\beta}$ from their posterior distribution and imputing the missing values Y_j by drawing from the conditional distribution $P(Y_j|\mathbf{Y}_{-j};\boldsymbol{\beta})$. This corresponds with a Gibbs sampling scheme if the joint distribution of the variables can be constructed from their univariate conditional distributions and if the distribution from which parameters are drawn can be constructed from the joint distribution of the variables and an appropriate prior distribution (Van Buuren et al., 2011). These two conditions are not fulfilled when using MICE with categorical data, which means that the algorithm is not an exact Gibbs sampler. MICE can be conducted using the R package MICE (Van Buuren et al., 2011) or the STATA (StataCorp, 2011) package ICE (Royston, 2009; White, Royston, & Wood, 2010).

2.2.2.3 *Other Incomplete-Data Methods*

We have three remarks on other incomplete-data methods. First, besides MLID and multiple imputation, there are two other categories of incomplete-data methods: the *fully Bayesian* method (Ibrahim, Chen, Lipsitz, & Herring, 2005), and *weighted estimating equations* (Allison, 2001). We did not consider these two approaches to limit the scope of this paper. A full Bayesian analysis with for example WinBugs is in fact similar to both MLID and multiple imputation; that is, the parameters of the substantive model are estimated using the incomplete data using an algorithm containing a step in which the missing values are imputed (Ibrahim et al., 2005). Results can be expected to be similar to MLID. Weighting is typically used to deal with completely missing data and has limited practical use with partially missing data (Kang & Schafer, 2007). It may moreover yield instable estimates in the presence of influential weights (Vansteelandt , Carpenter, and Kenward, 2010).

Second, a popular imputation model for multiple imputation is the multivariate normal distribution (Schafer & Graham, 2002). The method is robust against deviations from normality (Graham & Schafer, 1999), and may even perform well for categorical data (Bernaards, Belin, & Schafer, 2007), although some studies reported serious bias (Allison, 2005; Horton, Lipsitz, & Parzen, 2003). We did not consider incomplete data-methods that were not designed for categorical data as these methods are not suitable for nominal variables (e.g., blood type, eye color, surgical outcome).

Third, the best known ad hoc method is probably complete-case analysis, in which only the observations without any missing values are used to estimate the substantive model. In other words, subjects who have at least one missing value are discarded from the analysis. Hence, in contrast to MLID, complete-case analysis does not incorporate all available information. Complete-cases analysis reduces power and may yield biased parameter estimates for the substantive model if the data are not missing completely at random (MCAR; Little & Rubin, 2002); this MCAR assumption is considered to be unrealistic in most situations (Schafer & Graham, 2002). Complete-case analysis is included in Study 2 and the real-data example. For Study 2, the number of variables was too large for more preferable benchmarks such as MILL and MLID.

2.2.3 Advantages, Disadvantages, and Unresolved Issues of the Incomplete-Data Methods

2.2.3.1 *Practical issues*

For application of the incomplete-data methods, sample size, complexity of the association structure in the data, and percentage of missingness are not restrictive for any of the methods. A limitation of MILL is that it cannot handle large numbers of variables because the number of cells in the contingency table that has to be evaluated in the log-linear model becomes too large. For example, the number of cells that need be evaluated exceeds one million for 20 dichotomous variables and one billion for 30 dichotomous variables, 19 trichotomous variables, or 13 variables with five categories. In cases where the substantive model contains fewer variables than available in the dataset, a possible solution is to consider only those variables that are used in the substantive model. However, following Schafer's notion on the number of variables, using only a small number of variables for the imputation model may result in biased parameter estimates. For MILC and MICE, large numbers of variables do not pose a problem. A potential problem for MLID is that it usually requires specialized software, depending on the substantive model that one wants to estimate, whereas standard data-analysis techniques can be applied after the imputation phase of MILL, MILC and MICE. Moreover, MLID can only be used if the number of variables in substantive model is not too large.

2.2.3.2 *Bias*

We consider three possible causes of bias in the parameter estimates: First, non-ignorable missingness in the data. Following Schafer's notion on the number of variables, it is suggested that the inclusion of many variables in the imputation model makes it more likely that violations of ignorability are minor. The second possible cause of bias is misspecification of the imputation model so that it is too parsimonious. The imputation model should be as general as possible; this ensures that the imputed values behave as neutral as possible in subsequent analyses (Schafer, 1997). Hence, the main criterion of an adequate imputation model is whether it captures all the associations between categorical variables that exist on the population level (Schafer, 1997). The third possible cause of bias is misspecification of the substantive model. However, this is unrelated to the incomplete-data method being used and is not pursued further.

For MLID, no imputation model needs to be specified but a violation of the ignorability assumption may result in biased parameter estimates. Statistical analyses that are based on MLID only include those variables in the data that are substantively relevant,

possibly excluding many variables. When the number of variables in the substantive model is small, then, following Schafer's notion on the number of variables, the missingness mechanism in the reduced data is less likely to be ignorable. Simulation studies showed that under ignorable missingness, MLID yields unbiased parameter estimates (Schafer, 1997).

For MILL, the imputation model being too parsimonious is not an issue because the imputation model is typically the saturated model. However, MILL can handle only a limited number of categorical variables. As a result, following Schafer's notation on the number of variables, the missingness mechanism in the reduced data may not be treated as ignorable possibly resulting in biased parameter estimates. Simulation studies showed that under ignorable missingness, MILL yields unbiased parameter estimates (Schafer, 1997; Vermunt et al., 2008).

For MILC and MICE, the amount of non-ignorable missingness may be reduced if the data contain many variables relevant for predicting the missing values (Schafer's notion on the number of variables) because both methods can handle a very large number of (auxiliary) variables. For MICE, it is unknown which of the two variants yields the least bias, for MILC, it is expected that a large number of latent classes, MILC (Large), produces less bias than a smaller number of latent classes, MILC (AIC).

2.2.3.3 *Stability*

We consider three possible causes that influence the stability of parameter estimates in the presence of incomplete data. A first possible cause is a too small effective sample. It is well known that sample size has a positive effect on stability (Neyman & Pearson, 1933). None of the incomplete-data methods under investigation unduly reduce the effective sample size, in the way some ad hoc methods do (e.g., complete-case analysis, pair-wise deletion). However, it is unknown whether the incomplete-data methods under investigation yield the same stability of parameter estimates given a fixed sample size. A second possible cause is misspecification of the imputation model so that it is too complex. This is the well known tradeoff between bias and stability: If the imputation model is too parsimonious it may result in biased outcomes, if it is too complex, it may result in less stable outcomes. For most researchers unbiased parameter estimates are more important than stable parameter estimates. The third possible cause of instability is misspecification of the substantive model so that it is too complex. However, this is unrelated to the incomplete-data method being used and is not pursued further.

Only for the second possible cause of instability, an overly complex imputation model, we have expectations for the incomplete-data methods under investigation. MLID does not require an imputation model, so no loss of stability can ensue from an overly complex imputation model. For MILL, the imputation model is saturated meaning that it is expected to be overly complex in most cases. Therefore, a certain loss of stability is expected for MILL in comparison to MLID.

For MILC, the two variants are expected to differ in stability because their respective imputation models differ in complexity. MILC (Large) uses a relatively large number of latent classes which means that its imputation model is expected to be able to capture every possible association. As is the case with MILL, results produced by MILC (Large) are expected to lose a certain degree of stability because its imputation model is expected to be overly complex. MILC (AIC) estimates the required number of latent classes using AIC, which results in a relatively small number of latent classes. Therefore, its imputation model is more parsimonious and its results are expected to be more stable than MILC (Large).

For MICE, the two variants differ in the conditional imputation model that is used. The stability of MICE depends on the degree to which higher order associations are included in the conditional imputation model. The default setting of MICE (PMM) only includes main effects. However, because predictive mean matching is used, all associations can be picked up for datasets with a small number of variables. Therefore, we expect that the stability of the parameter estimates produced by MICE (PMM) is similar to MILL and MILC (Large). The default setting of MICE (LOG) also only includes main effects. Therefore, MICE (LOG) is expected to have relatively stable results.

2.2.3.4 *Bias of the Standard Errors*

It is unknown whether the six incomplete-data methods overestimate or underestimate the standard errors of parameter estimates. Hence, we have no specific expectations with regard to the bias of the standard errors.

2.3 Study 1: Bias, Stability, and Bias in Standard Errors Produced by MILC, MICE, MILL, and MLID for a Small Number of Dichotomous Variables.

In Study 1, we compared incomplete-data methods MILC (AIC), MILC (Large), MICE (PMM), and MICE (LOG) to MLID and MILL, on bias of the parameter estimates, stability of the parameter estimates, and bias of the standard errors. Because MLID and MILL can

handle only a limited number of variables, the number of variables was kept small. The design of Study 1 was motivated by the study of Kurian et al. (2010), who studied several predictors of a single outcome variable reduced "length of hospital stay" using logistic regression.

2.3.1 Method

2.3.1.1 *General setup*

The set up of the simulation study was as follows. First, we sampled complete datasets from a population model. Second, we created incomplete datasets by deleting variable scores according to an MAR missingness mechanism. Third, for each incomplete dataset we constructed five completed datasets using a missing-data method. Fourth, we used the completed datasets to estimate the parameters of the regression part of the population model and we reported the bias and stability of the parameter estimates.

The population model was defined for five dichotomous predictor variables $Y_1, \ldots, Y_5$, and one dichotomous outcome variable Y_6. The categories were coded 0 and 1 (dummy coding). Dummy coding was used because it is the most commonly used coding scheme for logistic regression models. The associations among the predictor variables $Y_1, \ldots, Y_5$ were described by log-linear model

$$\log P(Y_1, Y_2, Y_3, Y_4, Y_5) = -1.47 + \sum_{j=1}^{5} -2 \cdot Y_j + \sum_{j=1}^{4} \sum_{k=j+1}^{5} 1 \cdot Y_j Y_k \qquad (2.2)$$

Outcome variable Y_6 was related to the predictor variables by logit model

$$\text{logit}(Y_6) = \beta_0 + Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + Y_4 + Y_5 - \beta_{23} Y_2 Y_3, \qquad (2.3)$$

which contains main effects of the predictor variables as well as the interaction effect of Y_2 and Y_3. The strength of the interaction term, β_{23}, was manipulated in the study. The coefficients β_0, β_2 and β_3 are changed together with β_{23} so that the average logit and the average effects of Y_2 and Y_3 remain constant across conditions. Complete datasets were created by sampling from $P(Y_1, Y_2, Y_3, Y_4, Y_5)$ (Eq. 2.2), and $P(Y_6|Y_1, Y_2, Y_3, Y_4, Y_5)$ (Eq. 2.3).

Variables Y_1 and Y_2 had missing values that were MAR. Variables R_1 and R_2 indicated whether a score was missing, $R_i = 0$, or observed, $R_i = 1$, for Y_1 and Y_2, respectively. Missing values in Y_1 were created using logistic regression model

$$\text{logit}(R_1) = \gamma_1 + 1.09 \cdot Y_3 + 2.01 \cdot Y_4 - .79 \cdot Y_3 Y_4, \qquad (2.4)$$

and missing values in Y_2 were created using logistic regression model

$$\text{logit}(R_2) = \gamma_2 + 1.04 \cdot Y_5 + 1.94 \cdot Y_6 - .74 \cdot Y_5 Y_6, \qquad (2.5)$$

The total percentage of missingness (one of the predictor variables in Study 1, to be discussed later) was manipulated by changing the intercepts (γ_1 and γ_2) in eq. (2.4) and (2.5). This approach allows for varying the total percentage of missingness without altering the strength of associations between the predictor variables and the missingness indicator variable in eq. (2.4) and (2.5).

For each incomplete dataset, five completed datasets were created using a multiple imputation method and for each completed dataset logistic regression model

$$\text{logit}(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 + \beta_{23} Y_2 Y_3 \qquad (2.6)$$

was estimated. The simple rules introduced by Rubin (1987) were used to combine the five sets of regression parameter estimates. It should be noted that $m = 5$ completed datasets is usually considered to be sufficient to obtain stable results (Schafer, 1997). However, other researchers have argued that *m* should be based on the total percentage of missingness; for example, *m* should equal the total percentage of missing data to obtain a sufficient degree of stability in the results (White et al., 2010). In many cases, this would render *m* larger than five. We note that this is especially important for a single analysis; in a simulation study the size of *m* has much less influence because of the large number of replications.

Three software packages were used for multiple imputation and parameter estimation. Data were generated using software package LEM, methods MILC and MILL were conducted using the software program LatentGOLD 4.5, and for MICE we used the R package MICE. After multiple imputation, the substantive model, defined in Equation 2.6, was estimated using LatentGOLD 4.5. For MLID, the substantive model was estimated using LEM.

2.3.1.2 *Predictor variables*

Incomplete-data method was a within factor with six levels: MILC (AIC), MILC (Large), MICE (PMM), MICE (LOG), MLID, and MILL. The incomplete-data methods determine the imputation model and may thus affect both bias and stability.

Strength of the interaction term was a between factor with three levels that was manipulated by varying parameter β_{23} in eq. (2.3). The levels were: no three-way association ($\beta_{23} = .00$), medium ($\beta_{23} = -.80$), and strong ($\beta_{23} = -2.00$). Strength of the three-way association sets requirements for the complexity of the imputation model. If this effect increases, a more complex imputation model is required to pick up the associations in the data. It is expected that strength of the three-variable association affects both bias and stability.

Percentage of missingness was a between factor with three levels: moderate (10% missingness), high (20% missingness), and extreme (40% missingness). The percentage of missingness was manipulated by varying parameters γ_1 and γ_2 in eq. (2.4) and (2.5), respectively. For 10% missingness, $\gamma_1 = -2.46$ and $\gamma_2 = -2.53$, for 20% missingness, $\gamma_1 = -1.41$ and $\gamma_2 = -1.44$, and for 40% missingness, $\gamma_1 = -.39$ and $\gamma_2 = -.41$. As the percentage of missingness increases, the imputation model becomes more important. The condition with 40% of missingness is included because the consequences of an inadequate imputation model are magnified by an increase in the percentage of missingness.

Sample size was a between factor with two levels: Small (*N* = 200) and large (*N* = 1000). Sample size is expected to predominantly affect stability. In particular, the aim was to examine how sample size is related to the stability of the statistical results in the analysis of interest for each missing-data method.

The four predictor variables were fully crossed producing a $5 \times 3 \times 3 \times 2$ design, with 1,000 replications for each of the 18 combinations of the between-subjects variables.

2.3.1.3 *Outcome variables*

The outcome variables were bias of parameter estimates, standard deviation of parameter estimates across replications, and bias of the reported standard errors (Neyman & Pearson, 1933; Schafer & Graham, 2002). Let $\hat{\beta}_{bj}$ denote a parameter estimate of the *j*th variable (Eq. 2.6) in replication *b* (*b* = 1, …, 1,000), then the bias over 1,000 replications was computed as

$$\text{bias} = \frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{\beta}_{bj} - \beta_j\right)$$

Stability, denoted by $\text{sd}(\hat{\beta}_j)$, was measured by the standard deviation of parameter estimates across replications and was computed as

$$\text{sd}\left(\hat{\beta}_j\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_{bj} - \bar{\beta}_j\right)^2}.$$

Let $\text{se}\left(\hat{\beta}_{bj}\right)$ denote the estimated standard error of parameter estimate $\hat{\beta}_{bj}$. Bias of the reported standard errors (BSE) was computed as

$$\text{BSE} = \frac{1}{1000}\sum_{b=1}^{1000}\left[\text{se}\left(\hat{\beta}_{bj}\right) - \text{sd}\left(\hat{\beta}_{j}\right)\right],$$

The bias and stability of parameter estimates, and bias of the standard errors were only considered for parameters β_2 (a main effect that is influenced by the interaction effect β_{23}), β_4 (a main effect that is not influenced by the interaction effect β_{23}), and the interaction effect β_{23}.

2.3.2 Results

2.3.2.1 *Bias*

Table 2.1 shows the most important results for bias for 40% missingness. For lower percentages of missingness, the pattern of the bias was similar but the absolute values were smaller. This confirms that for larger percentages of missingness, the imputation model becomes more important. The most important result is that incomplete-data methods MLID, MILC (AIC) and MICE (LOG) produced large bias in the estimates of β_2 and β_{23} when there was an interaction effect in the data ($\beta_{23} \neq 0$), whereas estimates of β_4, a parameter not influenced by an interaction effect, showed much less bias. These results suggest that MILC (AIC) and MICE (LOG) have imputation models that are too parsimonious to pick up the three-way association in the data. Furthermore, the results suggest that MLID cannot handle very well the combination of a small sample size and a complex association. Seemingly, the asymptotic property of unbiased parameter estimates is not fully established under these circumstances. A second result is that MILL, which we used as a gold standard, produced similar bias or sometimes more bias (e.g., for β_2 in condition $\beta_{23} = -2,\ N = 200$) than MILC (Large) and MICE (PMM). A third result is that the bias was slightly larger for $N = 200$ than for $N = 1000$ which indicates that increased sampling variability may somewhat increase bias.

Table 2.1: *Bias in the Estimates of Three Logistic Regression Coefficients for Six Incomplete-Data Methods,Three Levels of Strength of the Three-Variable Association ($\beta_{23} = 0$, $\beta_{23} = -.8$, $\beta_{23} = -2$), Two Sample Sizes (200, 1000), and 40% Missingness. Remarkable Bias is Printed in Boldface.*

		Incomplete-data method					
N	RC	MLID	MILL	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
200	$\beta_2 = 1$	.040	.076	.082	.025	.051	.073
	$\beta_4 = 1$	.065	.067	.061	.041	.068	.061
	$\beta_{23} = 0$	.046	**-**.014	**-**.008	.031	.014	.009
	$\beta_2 = 1.4$	.050	.094	.086	.040	.058	.058
	$\beta_4 = 1$	.071	.057	.058	.042	.062	.057
	$\beta_{23} = -.8$	**-**.036	.010	.022	**.258**	.036	**.310**
	$\beta_2 = 2$	**.091**	**.103**	.057	**-.241**	.057	**-.357**
	$\beta_4 = 1$	.058	.040	.036	.002	.055	.021
	$\beta_{23} = -2$	**-.144**	-.074	-.006	**.541**	**-**.025	**.733**
1000	$\beta_2 = 1$	**-**.005	.010	.009	**-**.003	.011	.008
	$\beta_4 = 1$	.015	.013	.015	.012	.016	.014
	$\beta_{23} = 0$	**-**.019	**-**.004	**-**.006	**-**.001	**-**.006	**-**.004
	$\beta_2 = 1.4$	.026	.036	.036	**-.088**	.036	**-.136**
	$\beta_4 = 1$	.005	.003	.004	**-**.004	.003	**-**.002
	$\beta_{23} = -.8$	**-**.010	**-**.016	-.015	**.205**	**-**.014	**.302**
	$\beta_2 = 2$	.006	.033	.028	**-.216**	.031	**-.410**
	$\beta_4 = 1$	.014	.011	.011	**-**.015	.012	**-**.013
	$\beta_{23} = -2$	**-**.036	**-**.046	**-**.041	**.457**	**-**.056	**.765**

Note: N = sample size; β_{23} = strength of three-variable association; RC= regression coefficient. For MILC (AIC) the average number of classes indicated by AIC ranged from 2.8 ($N = 200$, $\beta_{23} = -2$) to 3.8 ($N = 1000$, $\beta_{23} = -2$), for MILC (Large) a constant number of 12 classes was used.

2.3.2.2 *Stability*

Table 2.2 shows the most important results for stability for 40% missingness. The most important result is that stability does not change dramatically across incomplete-data methods. MILC (AIC) was slightly more stable than MILC (Large), and MICE (LOG) is slightly more stable than MICE (PMM). This was expected because MILC (AIC) and MICE (LOG) are more parsimonious than MILC (Large) and MICE (PMM), respectively.

Table 2.2: *Stability of the Estimates of Three Logistic Regression Coefficients for Six Incomplete-Data Methods, Three Different Levels of Strength of the Three-Variable Association ($\beta_{23} = 0,\ \beta_{23} = -.8, \beta_{23} = -2$), Two Sample Sizes (200, 1000), and 40% Missingness* .

		Incomplete-data method					
N	RC	MLID	MILL	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
200	$\beta_2 = 1$	.862	.845	.858	.600	.951	.738
	$\beta_4 = 1$	.500	.496	.502	.445	.518	.501
	$\beta_{23} = 0$	1.19	1.17	1.17	.840	1.23	.729
	$\beta_2 = 1.4$	.938	.931	.924	.679	1.01	.780
	$\beta_4 = 1$	.505	.506	.510	.448	.533	.508
	$\beta_{23} = -.8$	1.21	1.22	1.20	.862	1.26	.737
	$\beta_2 = 2$	.916	.956	.948	.748	1.02	.806
	$\beta_4 = 1$	.494	.510	.515	.449	.537	.501
	$\beta_{23} = -2$	1.24	1.25	1.25	.917	1.29	.771
1000	$\beta_2 = 1$	.344	.373	.370	.264	.380	.301
	$\beta_4 = 1$	.203	.206	.206	.188	.206	.205
	$\beta_{23} = 0$	.479	.515	.509	.362	.522	.306
	$\beta_2 = 1.4$	.365	.389	.384	.291	.392	.313
	$\beta_4 = 1$	.205	.207	.208	.188	.207	.204
	$\beta_{23} = -.8$	.470	.507	.494	.376	.501	.310
	$\beta_2 = 2$	.377	.407	.404	.342	.402	.306
	$\beta_4 = 1$	.200	.205	.205	.185	.205	.197
	$\beta_{23} = -2$	.482	.526	.517	.451	.514	.298

Note: N = sample size; β_{23} = strength of three-variable association; RC= regression coefficient.

The expected result that MILL would be less stable than MLID was not demonstrated. As expected, sample size had a positive effect on stability. For small samples ($N = 200$) the stability could be considered low, resulting in low power. For example, the population value of β_4 was equal to 1, but in case $\text{sd}(\hat{\beta}_4) = .501$ (MILL, medium three-variable association), which is even one of the smaller standard deviations we found, one may expect to find estimates of β_4 between .02 and 1.98 (95% confidence interval). For large samples ($N = 1000$), the stability is much better. Percentage of missingness also had a negative effect on the stability. This can be expected because a larger percentage of missingness in fact means a reduction of the sample size.

2.3.2.3 *Bias of the standard errors*

Table 2.3 shows the most important results for bias of the standard errors for 40% missingness. Bias of the standard errors was smaller for $N = 1000$ than for $N = 200$.

Table 2.3: *Bias in the Standard Errors of the Estimates of Three Logistic Regression Coefficients for Six Incomplete-Data Methods, Three Different Levels of Strength of Three-Variable Associations ($\beta_{23} = 0$, $\beta_{23} = -.8$, $\beta_{23} = -2$), Two Sample Sizes (200, 1000), and 40% Missingness. Remarkable Bias is Printed in Boldface.*

		Incomplete-data method					
N	RC	MLID	MILL	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
200	$\beta_2 = 1$	-.040	-.044	-.044	**.151**	**-.202**	.072
	$\beta_4 = 1$	-.025	-.021	-.022	.015	-.052	-.021
	$\beta_{23} = 0$	-.092	**-.108**	-.077	**.174**	**-.219**	**.296**
	$\beta_2 = 1.4$	-.086	-.077	-.064	**.116**	**-.214**	.073
	$\beta_4 = 1$	-.027	-.017	-.018	.022	-.054	-.018
	$\beta_{23} = -.8$	-.092	-.077	-.062	**.194**	**-.209**	**.315**
	$\beta_2 = 2$	-.032	-.065	-.045	**.080**	**-.199**	.065
	$\beta_4 = 1$	-.010	-.015	-.017	.024	-.052	-.012
	$\beta_{23} = -2$	-.074	-.064	-.054	**.164**	**-.193**	**.292**
1000	$\beta_2 = 1$	-.001	-.015	-.011	**.065**	**-.060**	.038
	$\beta_4 = 1$	-.006	-.005	-.005	.008	-.010	-.005
	$\beta_{23} = 0$	-.015	-.030	-.025	**.088**	**-.086**	**.130**
	$\beta_2 = 1.4$	-.017	-.023	-.021	.041	-.067	.029
	$\beta_4 = 1$	-.007	-.006	-.007	.009	-.011	-.004
	$\beta_{23} = -.8$	-.005	-.015	-.005	**.070**	**-.059**	**.126**
	$\beta_2 = 2$	-.012	.040	-.036	**.080**	**-.074**	.032
	$\beta_4 = 1$	.002	-.002	-.000	.024	-.006	.001
	$\beta_{23} = -2$	.002	.028	-.028	**.164**	**-.068**	**.133**

Note: N = sample size; β_{23} = strength of three-variable association; RC= regression coefficient.

Bias of the standard errors was largest for the parameters associated with the three-variable association (β_2 and β_{23}). For $N = 200$, MILL and MILC (Large) had the smallest bias, whereas MILC (AIC) and MICE (LOG) overestimated the standard errors and MLID and MICE (PMM) underestimated the standard errors. For $N = 1000$, MILL, MILC (AIC), MILC (Large), and MICE (PMM) had the smallest bias, whereas MLID underestimated and MICE (LOG) overestimated the standard errors. This renders MILC (Large) as the most favorable incomplete-data method with respect to bias in standard errors.

2.4 Study 2: Bias, Stability, and Bias in Standard Errors Produced by MILC, MICE and Complete-Case Analysis for a Larger Number of Trichotomous Variables.

In Study 2, we compared incomplete-data methods MILC (AIC), MILC (Large, *K*=33), MICE (PMM), and MICE (LOG) to complete-case analysis, on bias of the parameter estimates, stability of the parameter estimates, and bias of the standard errors. Benchmarks MLID and MILL could no longer be used because the number of variables (11) was too large. The main question for Study 2 was whether MILC and MICE would also work for polytomous categorical data and for large numbers of possible response patterns. In Study 1, the number of possible response patterns was 2^6 = 64, whereas in Study 2, the number of possible response patterns was increased to 3^{11} = 177,147. The main objective for the design of Study 2 is that the associations among the variables need to be complex, to test whether the incomplete-data methods can pick up the associations correctly.

2.4.1 Method

2.4.1.1 *General set up*

In Study 2, the population model from which the complete datasets were sampled contained eight trichotomous predictor variables ($Y_1, \ldots, Y_8$) and three trichotomous outcome variables (Y_9, Y_{10}, and Y_{11}). The categories were coded 0, 1, and 2. The associations among the 11 variables are described by a path model for categorical data (Goodman, 1973) containing one-, two-, and three-way associations (see Figure 2.1 for a graphical representation and the Appendix for the chosen parameter values).

Variables Y_1, Y_3, Y_4 and Y_{11}had missing values; the other variables were completely observed. The missingness mechanism was MAR, and rather complex. For Y_1 and Y_3, the missingness depended on Y_2 and Y_9. Let R indicate whether (score 1) or not (score 0) a score is observed. Both for Y_1 and Y_3, the logit of R was $\text{logit}(R) = -5.06 - 2 \cdot Y_9 + 3 \cdot Y_2$, resulting in approximately 20% missing values for each variable. Similarly, for Y_4 and Y_{11}, the missingness depended on Y_7 and Y_9. Both for Y_4 and Y_{11}, $\text{logit}(R) = -5.50 + 3 \cdot Y_9 - 1.5 \cdot Y_7$, also resulting in approximately 20% missing values for each variable. This procedure kept the total percentage of missingness constant at $4/11 \cdot 20\% + 7/11 \cdot 0\% = 7.27\%$.

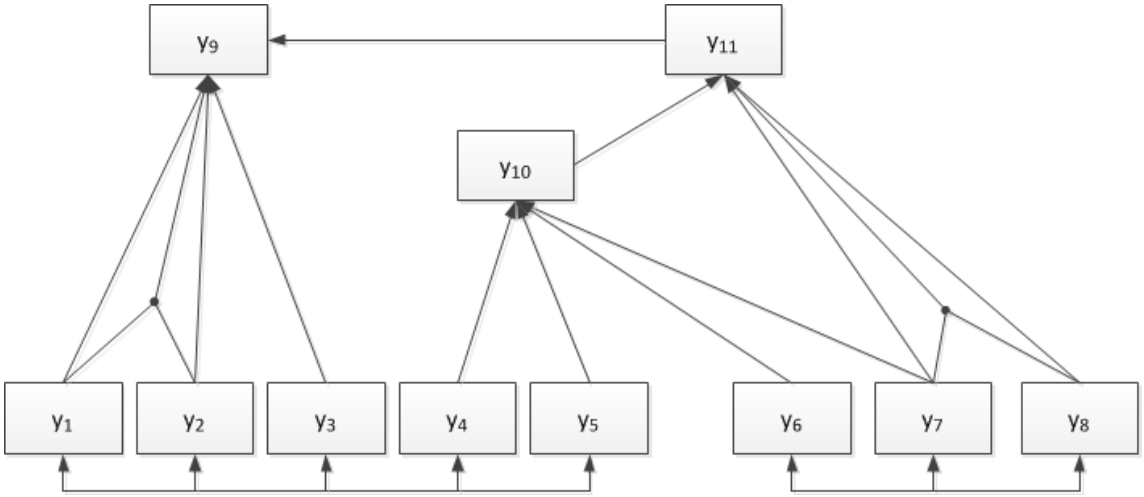

Figure 2.1. Population model of the second Study. The model contains 11 trichotomous variables: Y_1 through Y_8 are predictor variables, and Y_9 through Y_{11} are outcome variables.

For each incomplete dataset, the multiple imputation methods created $m = 5$ completed datasets. For complete-case analysis, a complete dataset was obtained by simply deleting every row that contained at least one missing value.

The substantive model was an adjacent category ordinal logit model (Agresti, 1990) for outcome variable Y_9 containing Y_1, Y_3, Y_4, and Y_{11} as predictors. The logit equation has the form

$$\text{logit}(Y_9 = j | Y_9 = j-1 \text{ or } Y_9 = j) = \beta_{0j} + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_{12} Y_1 Y_2,$$

for $j = 2, 3$. Note that the substantive model is part of the population model (Figure 2.1) and includes the main effects of the predictors of Y_9, and the interaction effect of Y_1 and Y_2. The latter implies a three-variable association among Y_1, Y_2, and Y_9

Three software packages were used for data generation, incomplete-data handling, and estimating the substantive model. Complete and incomplete data were generated by LEM. The imputation phase of MILC (Large) and MICE (PMM) was performed using LatentGOLD and the R package MICE, respectively. LatentGOLD was used to estimate the substantive model for by MILC (Large) and MICE (PMM), using the completed datasets, and for MLID and complete-case analysis.

2.4.1.2 *Design*

We only varied sample size and incomplete-data method. Sample size had two levels: medium ($N = 500$) and large ($N = 1000$); incomplete-data method had five levels: MILC (AIC), MILC (Large), MICE (LOG), MICE (PMM), and complete-case analysis. This yields

a 5×2 design. The outcome variables were equivalent to those in Study 1 (bias, stability, and bias of standard errors).

2.4.2 Results

2.4.2.1 *Bias*

Table 2.4 shows the bias for β_2, the main effect of a predictor that is also involved in the interaction effect; β_3, the main effect of a predictor not involved in the interaction effect; and β_{12}, the interaction effect itself. The most important result is that MICE (PMM) and MICE (LOG) produced relatively large bias in the estimates of β_3 and β_{12}, suggesting that the imputation models of MICE (LOG) and MICE (PMM) do not correctly pick up the three-way association in the data. Furthermore, complete-case analysis produced very large bias in the estimates of β_2 and β_{12}, confirming that complete-case analysis leads to biased results when the data are MAR. MILC (AIC) and MILC (Large) had a similar performance in terms of bias.

Table 2.4: *Bias in the Estimates of Three Logistic Regression Coefficients for Five Incomplete-Data Methods, Two Sample Sizes (500, 1000), and 20% Missingness on Four Variables. Remarkable Bias is Printed in Boldface.*

		Incomplete-data method				
N	RC	CC	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
500	$\beta_2 = -.45$	**.504**	**-**.033	**-**.028	**-**.035	**-**.036
	$\beta_3 = .5$	.017	.001	**-**.002	**.054**	**.051**
	$\beta_{12} = .45$	**-.110**	**-**.068	**-**.066	**-.114**	**-.113**
1000	$\beta_2 = -.45$	**.501**	**-**.027	**-**.025	**-**.033	**-**.035
	$\beta_3 = .5$	.019	.001	**-**.003	**.053**	**.046**
	$\beta_{12} = .45$	**-.113**	**-**.061	**-**.061	**-.116**	**-.114**

Note: N = sample size; RC = regression coefficient.

2.4.2.2 *Stability*

Table 2.5 shows the stability of β_2, β_3, and β_{12}. The most important result is that (almost) unbiased parameter estimates (see Table 2.4) showed similar stability across methods, whereas biased parameter estimates tended to be either more stable or more unstable.

Table 2.5: *Stability of the Estimates of Three Logistic Regression Coefficients for Five Incomplete-Data Methods, Two Sample Sizes (500, 1000), and 20% Missingness on Four Variables.*

		Incomplete-data method				
N	RC	CC	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
500	$\beta_2 = -.45$	.101	.290	.288	.293	.292
	$\beta_3 = .5$	.215	.218	.215	.240	.240
	$\beta_{12} = .45$	.146	.159	.157	.127	.130
1000	$\beta_2 = -.45$	.069	.272	.271	.275	.275
	$\beta_3 = .5$	.224	.222	.220	.247	.246
	$\beta_{12} = .45$	.134	.162	.161	.134	.137

Note: N = sample size; RC = regression coefficient.

This effect for was clearer for $N = 500$ than for $N = 1000$. For example, for the estimate of β_2, MILC (Large), MILC (AIC), MICE (PMM), and MICE (LOG) show similar bias and similar stability of parameter estimates. However, complete-case analysis overestimated β_2 and this estimate was too stable, whereas MICE (PMM) and MICE (LOG) overestimated β_3 and this estimate was too unstable.

2.4.2.3 *Bias of the standard errors*

Table 2.6 shows the bias of the standard errors of β_2, β_3, and β_{12}. Bias of the standard errors was smaller for *N* = 1,000 than for *N* = 500. The multiple imputation methods yielded similar upward bias in their standard errors, whereas complete-cases analysis tended to yield a larger overestimation of the standard errors. Bias was largest for the standard error of the interaction effect (β_{12}).

Table 2.6: *Bias in the Standard Errors of the Estimates of Three Logistic Regression Coefficients for Five Incomplete-Data Methods, Two Sample Sizes (500, 1000), and 20% Missingness on Four Variables.*

		Incomplete-data method				
N	RC	CC	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
500	$\beta_2 = -.45$	**.126**	.080	.080	.082	.081
	$\beta_3 = .5$	**.120**	.081	.083	.088	.084
	$\beta_{12} = .45$	**.156**	.100	.102	.106	.102
1000	$\beta_2 = -.45$	**.089**	.058	.058	.058	.057
	$\beta_3 = .5$	**.086**	.061	.062	.062	.059
	$\beta_{12} = .45$	**.111**	.076	.077	.075	.072

Note: N = sample size; RC = regression coefficient.

2.5 Real-data Example

We applied the most promising variants of MILC and MICE (i.e., MILC (Large, $K = 12$) and MICE (PMM)), complete-case analysis, and MLID to data from the Investigators of Projective Services Project for Older Persons (Blenkner, Bloom, & Weber, 1974), which have been discussed and analyzed earlier by Fuchs (Fuchs, 1982) to illustrate the MLID approach. The dataset contains the scores of 164 patients on six dichotomous variables (Table 2.7). One patient had a missing value on the physical status, 33 had a missing value for mental status, and 29 respondents had a missing value on both physical and mental status.

Table 2.7: Information on the Variables of the Protective Services Project for Older Persons.

Variable	Levels	Code
Mental status	poor, good	Y_1
Physical status	poor, good	Y_2
Age	less than 75, over 75	Y_3
Group membership	experiment, control	Y_4
Sex	male, female	Y_5
Survival status	deceased, survived	Y_6

The question of interest is whether the unexpected negative association between treatment and survival disappears when controlling for age, gender, physical status, and mental status. The substantive model predicts survival by the main effects of all variables, plus the interaction effect of mental status, Y_1, and physical status, Y_2. We defined the following regression model,

$$\text{logit}(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 + \beta_{12} Y_1 Y_2. \quad (2.7)$$

Contrary to Fuchs, we chose to include the interaction between mental status and physical status because we were interested in whether the imputation methods yielded similar results to MLID in a model containing a higher-order association. Once the data had been imputed using MILC (Large) and MICE (PMM), the substantive model defined in equation 2.7 was estimated. We also estimated the logistic regression model using MLID (as a benchmark) and complete-case analysis. Schafer's notion on the number of variables is of no concern in this analysis because the substantive model and the imputation model are identical; both models include all available variables. Therefore, the performance of MILC (Large), MICE (PMM), and complete-case analysis was assessed by comparing them to MLID (Table 2.8). For all incomplete-data methods, only age (β_3, negative effect) had a significant effect on survival status. The fact that all other effects were not statistically significant may be due to the small sample size. Nevertheless, it remains interesting to compare the parameter estimates across incomplete-data methods.

Table 2.8: *Estimated Logistic Regression Coefficients using MLID, Complete-Case Analysis, MILC (Large), and MICE (PMM).*

RC	Incomplete-data method							
	MLID		Complete-case		MILC (Large)		MICE (PMM)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_1	3.175	2.188	2.844	2.070	3.777	1.984	**2.271**	1.815
β_2	2.614	2.240	**1.816**	2.180	3.162	2.102	**1.439**	1.979
β_3	-1.417	.431	-1.568	.526	-1.380	.422	-1.426	.434
β_4	.459	.394	**.176**	.481	.496	.392	.475	.384
β_5	-.506	.417	**-.281**	.525	-.420	.437	-.583	.410
β_{12}	-1.017	1.269	**-.629**	1.217	-1.283	1.168	**-.382**	1.090

Table 2.8 shows that the estimates yielded by MILC (Large) were very similar to MLID, for all parameters. MICE (PMM) produced estimates of β_3, β_4, and β_5 that were very similar to MLID, but yielded relatively large differences for parameters β_1, β_2, and β_{12}. Complete-case analysis produced rather large differences for the estimates of β_2, β_4, β_5, and β_{12}. MLID, MILC (Large), and MICE (PMM) did not have large differences in the estimated standard errors. However, complete-case analysis yielded relatively large standard errors for parameters β_3, β_4, and β_5, compared to MLID.

2.6 Discussion

The aim of this paper was to investigate which incomplete-data method for categorical data should be recommended to practitioners. We assessed the performance of MILC and MICE with regard to three criteria, relative to MLID, MILL, and complete-case analysis. Based on the theoretical discussion, Study 1, and Study 2, MILC (Large) appears to be the incomplete-data method that meets the three criteria to the greatest extent. The other incomplete-data methods have one or more features that make them suboptimal. MILL cannot handle more than a few variables, MLID does not allow for the use of small substantive model as it can affect the MAR assumption, and may yield biased parameter estimates for a complex association in case of a small sample size. While in Study 1 MICE (PMM) performed rather well, Study 2 showed that MICE (PMM) may yield biased parameter estimates when the number of possible data pattern is large, especially when the sample size is small; the real data example also showed that MICE (PMM) cannot estimate a complex association very well. MILC (AIC) and MICE (LOG) may fail to capture higher-order associations in the data, which yields parameter estimates with an unacceptably high bias. In Study 2 it was demonstrated that complete-case analysis yields very large bias in the parameter estimates, and a loss of power due to inflated standard errors. The findings in the real-data example were consistent with these results.

A remaining issue with MILC is that there is not yet a guideline indicating the minimum number of required latent classes. The simulation study showed that overfit does not seem to be a problem, which was also argued by Vermunt et al. (2008), so one can always resort to estimating a latent class model with many classes. However, having a minimum of required latent classes would greatly facilitate the use of MILC because estimating latent class models with 40, 50, or 60 latent classes can be very time consuming. We showed that in case of a small table AIC is not a good criterion because the number of classes is too low; for a

large table MILC (AIC) yielded good results. A heuristic rule may be to use as many classes as there are categories in the data. For example, for a dataset consisting of 10 variables with three response categories and 5 dichotomous variables, the number of latent classes would be $10 \times 3 + 5 \times 2 = 40$ latent classes. Whether this heuristic rule is reasonable should be investigated in future research.

An additional comment should be made for MICE (LOG), as it may have been presented too negatively. The problem of MICE (LOG) is that in the default setting, interaction effects are not included in the conditional models. As a result the imputation model may be too parsimonious yielding biased parameter estimates. Further research would be required to investigate whether this method is able to produce unbiased results if the conditional model included higher-order interactions.

Lastly, we note that each incomplete-data method had to be applied using a different software package, as there is no package available that applies all of the methods. Further research is warranted to investigate the potential differences between implementations of MILC and MICE across software packages.

Appendix

Tables A1, A2, and A3 show the parameter values describing the population model in Study 2 (Chapter 2; Figure 2.1). All variables had three nominal response categories. Because dummy coding was used, the effect of the first category was zero (not displayed). For all two-way and three-way interactions only a single value is shown because the associations were defined to be ordinal (linear-by-linear). Table A1 shows the log-linear parameters describing $P(Y_1, Y_2, Y_3, Y_4, Y_5)$, Table A2 shows the log-linear parameters describing $P(Y_6, Y_7, Y_8)$, and Table A3 shows the logistic regression parameters relating predictor variables $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7$, and Y_8 to outcome variables Y_9, Y_{10}, and Y_{11}.

Table A1: *Log-linear Parameters Describing* $P(Y_1, Y_2, Y_3, Y_4, Y_5)$

$Y_1 = (-.05, .10)$	$Y_1, Y_2 = .30$	$Y_2Y_4 = .55$	$Y_1Y_4Y_5 = -.35$
$Y_2 = (-.10, .15)$	$Y_1, Y_3 = -.40$	$Y_2Y_5 = -.36$	$Y_2Y_4Y_5 = -.25$
$Y_3 = (.10, -.05)$	$Y_1, Y_4 = -.20$	$Y_3Y_4 = -.15$	$Y_1Y_2Y_3 = .55$
$Y_4 = (-.10, -.05)$	$Y_1, Y_5 = .50$	$Y_3Y_5 = -.05$	
$Y_5 = (.10, -.15)$	$Y_2, Y_3 = -.30$	$Y_4Y_5 = .30$	

Table A2: *Log-linear Parameters Describing* $P(Y_6, Y_7, Y_8)$,

$Y_6 = (-.30, -.15)$	$Y_6Y_7 = .32$
$Y_7 = (-.50, -.25)$	$Y_6Y_8 = -.40$
$Y_8 = (-.20, -.10)$	$Y_7Y_8 = .24$
	$Y_6Y_7Y_8 = .40$

Table A3: *Logistic Regression Parameters*,

$Y_9\|Y_1 = -.30$	$Y_{10}\|Y_4 = .22$	$Y_{11}\|Y_{10} = -.15$
$Y_9\|Y_2 = -.45$	$Y_{10}\|Y_5 = .32$	$Y_{11}\|Y_6 = -.30$
$Y_9\|Y_3 = .5$	$Y_{10}\|Y_6 = .42$	$Y_{11}\|Y_7 = .35$
$Y_9\|Y_1Y_2 = .45$	$Y_{10}\|Y_7 = -.38$	$Y_{11}\|Y_8 = .10$
$Y_9\|Y_{11} = .35$	$Y_{10}\|Y_8 = .34$	$Y_{11}\|Y_6Y_7 = .40$
	$Y_{10}\|Y_6Y_7 = -.14$	

Chapter 3

Divisive Latent Class Modeling as a Density Estimation Tool for Categorical Data

Abstract

Traditionally latent class (LC) analysis is used by applied researchers as a tool for identifying substantively meaningful clusters. More recently, LC models have also been used as a density estimation tool for categorical variables. We introduce a divisive LC (DLC) model as a density estimation tool that may offer several advantages in comparison to a standard LC model. When using an LC model for density estimation, a considerable number of increasingly large LC models may have to be estimated before sufficient model-fit is achieved. A DLC model consists of a sequence of small LC models. Therefore, a DLC model can be estimated much faster and can easily utilize multiple processor cores, meaning that this model is more widely applicable and practical. In this study we describe the algorithm of fitting a DLC model, and discuss the various settings that indirectly influence the precision of a DLC model as a density estimation tool. These settings are illustrated using a synthetic data example, and the best performing algorithm is applied to a real-data example. The generated data example showed that, using specific decision rules, a DLC model is able to correctly model complex association among categorical variables.

This chapter was submitted for publication.

3.1 Introduction

Traditionally, latent class (LC) analysis (Lazarsfeld, 1950; also see, e.g., Collins & Lanza, 2010; Goodman, 1974; Hagenaars & McCutcheon, 2002; Magidson & Vermunt, 2004; McCutcheon, 1987; Rindskopf & Rindskopf, 1986) is used as a statistical method to identify substantively meaningful groups from multivariate categorical data. For example, Keel et al. (2004) distinguished 4 LCs of people with eating disorders that were labeled 'restricting anorexia nervosa', 'anorexia nervosa and bulimia nervosa with the use of multiple methods of purging', 'restricting anorexia nervosa without obsessive-compulsive features', and 'bulimia nervosa with self-induced vomiting as the sole form of purging'. To facilitate interpretation, it is desirable to keep the number of LCs small, and because the interpretation of the LCs is based on the estimated model parameters, it is also desirable that the LC model is identifiable (e.g., Goodman, 1974) and the global maximum of the likelihood has been found.

More recently, LC models have been used in a different way: as estimators of the joint density of a set of categorical variables. The often complex multivariate density is approximated by a finite mixture of simpler multinomial densities. For example, density estimation by means of an LC model has been used for multiple imputation of categorical data (Gebregziabher & DeSantis, 2010; Van der Palm, Van der Ark, & Vermunt, 2013a; Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008), for smoothing large sparse contingency tables (Linzer, 2011), for estimating test-score reliability (Van der Ark, Van der Palm, & Sijtsma, 2011), and for summarizing image-data bases for pattern recognition (Bouguila & ElGuebaly, 2009). The idea of approximating a complex density by a mixture of simpler densities is well-known in finite mixture modeling (e.g., McLachlan & Peel, 2000, pp. 11-14) but the majority of research has focused on mixtures of continuous distributions (e.g., Everitt, Landau, & Leese, 2001, pp. 8-10). The most important issue when using LC models to estimate densities is the precision of the density estimate. Depending on the application of interest, the two-way, three-way, or higher-way interactions among the variables should be accurately described by the LC model. In this context, the LC model is solely used as a tool, and the substantive interpretation of the LCs is not important. Consequently, for density estimation, issues such as model identification, convergence to the global maximum, and having as few LCs as possible do not play a dominant role.

For datasets containing a large number of variables, density estimation by means of an LC model is problematic because a large number of LCs is usually required for precise density estimation. Let LC(K) denote an LC model with K classes. For example, Vermunt et

al., (2008) used AIC (Akaike, 1974) as a criterion and selected an LC(50) model to model a survey dataset of 79 variables. They indicated that even more LCs may have been needed for precise density estimation. A typical model-fit strategy is to estimate an LC(5), LC(10), LC(15), LC(20) model, etcetera, until the model fit no longer improves. This can be a very time-consuming process: For example, we reanalyzed the survey dataset used by Vermunt et al. (2008), containing 4292 cases and 79 categorical variables, and estimated an LC(5), LC(10), LC(15),…, LC(60), and LC(65) model. The analysis took 8 hours and 18 minutes (details in Table 3.1) on a, for current standards, very fast personal computer (i7 2600 quadcore processor, 8GB of internal memory). As parsimony with respect to the number of LCs is less important than precision of the parameter estimates, an LC(65) model may be taken as the final solution, as the LC(65) model yields the first considerable increase in AIC (Table 3.1). The long computation time and comparison of many LC models can be an obstacle for researchers, especially when a density has to be estimated multiple times (e.g., in multiple imputation based on bootstrap replications).

Table 3.1: *AIC and Computation Time for 13 LC Models Fitted on the ATLAS Survey Data*

Number of LCs	AIC	Computation time
5	469,422.728	0:04h
10	461,707.592	0:09h
15	458,369.833	0:14h
20	455,712.563	0:19h
25	453,578.838	0:25h
30	452,359.392	0:33h
35	451,258.196	0:39h
40	451,400.102	0:42h
45	449,592.844	0:54h
50	450,231.146	0:56h
55	449,465.083	1:04h
60	448,905.116	1:07h
65	456,395.814	1:12h

As a solution, we introduce the divisive LC (DLC) model as a fast alternative to the LC model for density estimation. First, we provide an intuitive description of the DLC model. Second, we discuss estimation of the DLC model and some arbitrary choices that can be made in the estimation algorithm. Third, using a generated data example, we compare the effect of

these different choices on the precision of estimating complex densities. Fourth, the best performing estimation algorithm is applied to a dataset that was also analyzed by Vermunt et al. (2008) using a standard LC model and we compare the results.

3.2 Divisive Latent Class Model

The DLC model is a top-down clustering of respondents into LCs. It is obtained by fitting a sequence of LC(1) and LC(2) models. Figure 3.1 shows a graphic representation of the structure of a DLC model. It has different levels. In general, let r denote the level in the sequential structure ($r = 0, 1, 2, ...$). At Level r, let $X^{(r)}$ denote the discrete latent variable with $Q^{(r)}$ LCs indexed by q. First, at Level 0, we start with an LC(1) model. Then, a decision is made whether or not the goodness of fit would be improved if the LC is split into two LCs. If an LC(2) model fits better than an LC(1) model, then we have two LCs at Level 1, otherwise we have one LC at Level 1 and the procedure stops. Suppose that at Level 1, we have two LCs (case depicted in Figure 3.1), then for each LC a decision is made whether or not the goodness of fit would be improved by splitting the LC again into two LCs. In Figure 3.1, the first LC is split whereas the second LC is not, yielding three LCs at Level 2. Once it has been decided that splitting an LC does not improve the goodness of fit, the particular LC remains unchanged for the rest of the procedure. The splitting procedure continues until splitting LCs no longer improves the goodness of fit. In Figure 3.1, this is the case at Level 5, where we have six LCs. Numbering the LCs per level from 1 to $Q^{(r)}$ is arbitrary. We used the following procedure: Once all LCs at Level r have been either split or maintained, Level r+1 has been established, and the LCs at Level r+1 are simply numbered from 1 tot $Q^{(r)}$ (from left to right in Figure 3.1). The DLC model is somewhat similar to divisive clustering, from which we took its name. The difference is that in a DLC model each respondent, at each level, has a probability to belong to each LC (soft partitioning), and in divisive clustering each respondent, at each level, belongs to a cluster with certainty (hard partitioning). The DLC model was inspired by the work of Ueda and Nakano (2000) and Wang, Luo, Zhang, and Wei (2004). Ueda and Nakano introduced a split-and-merge approach to estimating mixture models to overcome the problem of local maxima, whereas Wang et al. used a stepwise split-and-merge approach to determine the number of components of a mixture model. In a Bayesian framework, Hoijtink and Notenboom (2004; also, see Van Hattum & Hoijtink, 2009) used a stepwise approach to find the solution with the maximum number of LCs.

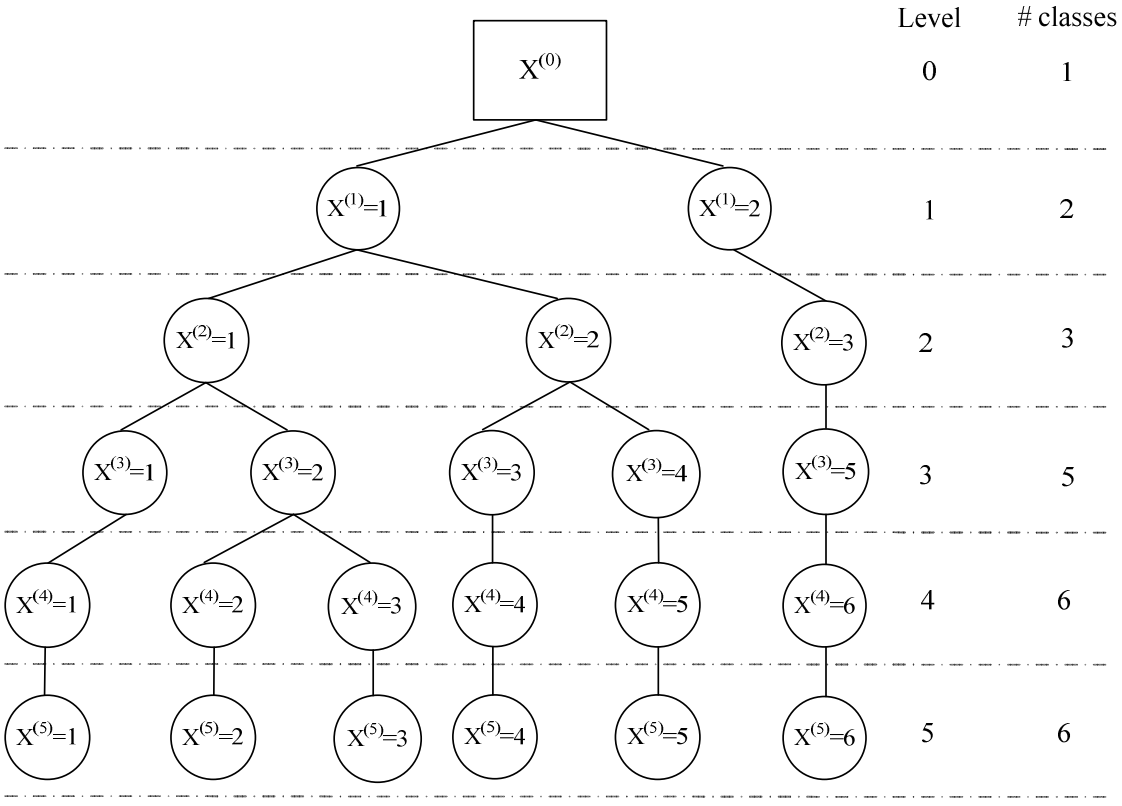

Figure 3.1: A graphic representation illustrating the DLC model. The sequence starts at Level 0, where the whole sample belongs to a single LC. In the subsequent levels an LC is split if that yields better fit. At Level 4, no classes have been split, rendering the LC model at Level 5 the final model.

The computational advantage of the DLC model over the standard LC model is that the estimation problem is broken down into a series of small problems coined *local problems*. Each local problem concerns the question whether splitting LC q at Level r will improve model fit (Figure 3.2). To this end, at Level $r + 1$, we estimate an LC*(1) model and an LC*(2) model. The asterisks indicate that the models are fitted on the *fuzzy* subsample in LC q at Level r rather than one the entire sample. If the LC*(2) model has a sufficiently better fit than the LC*(1) model, then LC q at Level r will be split. The estimation of the LC*(1) model and the LC*(2) model does not affect the LCs that are not part of the local problem. In the local problem, we arbitrarily number the LCs 1 and 2 for the LC*(2) model and 1 for the LC*(1) model. Note that an LC*(1) model and an LC*(2) model are estimated repeatedly—once for every local problem—in order to investigate whether a split is necessary.

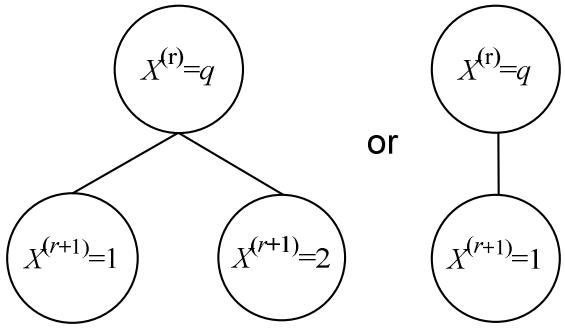

Figure 3.2: Graphic representation of a local problem: Should an LC at level r be split into two LCs at level r+1.

Let $\mathbf{y}_i = (y_{i1}, \dots, y_{ij}, \dots, y_{iJ})$ be the response vector of respondent i to manifest variables $Y_1, \dots, Y_j, \dots, Y_J$. In a standard LC($K$) model, the density $P(\mathbf{y}_i)$ is modeled as

$$\mathrm{P}(\boldsymbol{y_i}; \boldsymbol{\theta}) = \sum_{s=1}^{K} P(X = s) \prod_{j=1}^{J} P\big(y_{ij} \big| X = s\big). \tag{3.1}$$

The set of parameters, denoted by $\boldsymbol{\theta}$, consists of probabilities $P(X = q)$ —the probability that a randomly selected respondent belongs to LC q— and probabilities $P(y_{ij}|X = q)$ —the probability that a member of LC q has response y_{ij}. The log-likelihood for the LC(K) model is

$$\log L(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{N} w_i \log \sum_{s=1}^{K} P(X = s) \prod_{j=1}^{J} P\big(y_{ij} \big| X = s\big),$$

where w_i denotes the contribution of the response vector of respondent i to the log-likelihood. For standard LC models, the weights w_i are equal to 1 by definition.

For a local problem, depicted in Figure 3.2, an LC*(1) model ($K = 1$) and an LC*(2) model ($K = 2$) are estimated for the subsample in LC q at Level r rather than the entire sample. Hence, $P^*(\mathbf{y}_i) \equiv P(\mathbf{y}_i|X^r = q)$ is modeled rather than $P(\mathbf{y}_i)$. The LC model in Equation 3.1, then becomes

$$P^* \ (\mathbf{y}_i; \boldsymbol{\theta}^*) = \sum_{s=1}^{K} P^*\big(X^{(r+1)} = s\big) \prod_{j=1}^{J} P^*\big(y_{ij} \big| X^{(r+1)} = s\big). \tag{3.2}$$

In Equation 3.2, density $P^*(\mathbf{y}_i)$ is modeled by *local parameters*; the local parameters have the same interpretation as the parameters of a standard LC model, except for the fact that they are conditional on being member of LC $X^{(r)} = q$. Thus, $P^*\big(X^{(r+1)} = s\big) \equiv$

$P\left(X^{(r+1)}=s\middle|X^{(r)}=q\right)$ and $P^{*}\left(y_{ij}\middle|X^{(r+1)}=s\right)\equiv P\left(y_{ij}\middle|X^{(r+1)}=s;X^{(r)}=q\right)$. The local parameters are denoted by $\boldsymbol{\theta}^{*}$. The subsample in LC q at Level r is fuzzy because each respondent has a probability of belonging to this LC. The probability that a respondent having response vector $\mathbf{y}_i$ belongs to LC q at Level r is denoted as $P(X^{(r)}=q|\mathbf{y}_i)$, and referred to as the *posterior membership probability*. The posterior membership probability determines the weight of a particular respondent in the log-likelihood: $w_{iq}^{(r)}=\ P(X^{(r)}=q|\mathbf{y}_i)$. Hence, the log-likelihood for the LC*(K) model ($K=1,2$) in the local problem is

$$\log L(\boldsymbol{\theta}^{*};\mathbf{y})=\sum_{i=1}^{N}w_{iq}^{(r)}\log\sum_{s=1}^{K}P^{*}\left(X^{(r+1)}=s\right)\prod_{j=1}^{J}P^{*}\left(y_{ij}\middle|X^{(r+1)}=s\right). \qquad (3.3)$$

The parameter estimates of the selected LC model at Level r+1 in the local problem also yields a *local posterior membership probability* for $X^{(r+1)}=s$:

$$P^{*}(X^{(r+1)}=s|\mathbf{y}_i)\equiv P^{*}(X^{(r+1)}=s\big|\mathbf{y}_i;X^{(r)}=q\big)$$

$$=\frac{P^{*}(X^{(r+1)}=s)\prod_{j=1}^{J}P^{*}\left(y_{ij}\middle|X^{(r+1)}=s\right)}{\sum_{k=1}^{K}P^{*}(X^{(r+1)}=k)\prod_{j=1}^{J}P^{*}\left(y_{ij}\middle|X^{(r+1)}=k\right)}. \qquad (3.4)$$

For the LC*(2) model, the local posterior is the probability that a respondent belongs to each of the two LCs at level *r*+1, conditional on being member of LC q at level *r*. For the LC*(1) model, the local posterior equals 1 by definition. The local posterior membership probability is used to determine the weights of the respondents in the likelihood for the local problems at the next level. The weights at Level r+1 are obtained by multiplying the local posterior probability at Level $r+1$ and the (global) posterior probability at Level *r*:

$$w_{is}^{(r+1)}=P\left(X^{(r+1)}=s|\,\mathbf{y}_i\right)=P^{*}\left(X^{(r+1)}=s|\,\mathbf{y}_i\right)\times P\left(X^{(r)}=q|\,\mathbf{y}_i\right) \qquad (3.5)$$

The DLC model is estimated by the following iterative procedure.

1. *Initial step*: At Level 0, set $w_{i1}^{(0)}:=1$, $P\left(X^{(0)}=1\right):=1$, and $Q^{(0)}:=1$.
2. *Solve the local problem of LC q at Level r:* Estimate an LC*(1) and LC*(2) model for the fuzzy sample in LC $X^{(r)}=q$ by optimizing the likelihood in Equation 3.3, and choose either an LC*(1) model or LC*(2) model. If an LC*(1) model is chosen, LC $X^{(r)}=q$ is no longer considered for division at later levels and steps 3, 4, and 5 are skipped; see discussion hereunder.
3. *Compute the local posterior membership probabilities* (Equation 3.4).

4. *Update the posterior membership probabilities* from the local membership probabilities (Equation 3.5). The updated posterior membership probabilities are the weights for the local problems at Level $r + 1$.
5. *Update the parameter estimates from the posterior membership probabilities and local parameter estimates.*
   - $P\left(X^{(r+1)} = s\right) = \frac{1}{N}\sum_{i=1}^{N} P\left(X^{(r+1)} = s|\boldsymbol{y}_i\right)$
   - $P\left(y_{ij}\middle|X^{(r+1)} = s\right) = P^{*}\left(y_{ij}\middle|X^{(r+1)} = s\right)$
6. Repeat steps 2 through 5 for all LCs at Level r.
7. Renumber the LCs from 1 to $Q^{(r)}$, and let $r = r + 1$.
8. Repeat steps 2 to 7 until no more classes are split.

The remaining problem of DLC estimation is the choice of either the LC*(1) model or the LC*(2) model in each local problem (Figure 3.2). The choice depends on the required precision and the sample size. If the number of LCs becomes too large, the parameter estimates are unstable and the density estimate may be based on chance capitalization. If the number of LCs becomes too small, the density estimation may not be precise enough. Relevant factors for the choice of either the LC*(1) model or the LC*(2) model may be the difference in log-likelihood, the sample sizes in the LCs, and the size of residual associations between variables. In the generated data study, we investigate this issue.

3.3 Generated Data Study

The main question was whether a DLC model can precisely estimate a complex density that was not generated by an LC model. To this end, we used a DLC model to estimate a complex density under ideal circumstances, so removing all influences of sampling error. Additionally, we investigated different choices for selecting an LC*(1) model or an LC*(2) model in the local problem.

Method

We defined a complex population model, depicted in Figure 3.3, for 11 dichotomous variables ($Y_1, \ldots, Y_{11}$). The population model consists of two sets of independent variables ($\{Y_1, Y_2, Y_3\}$.and$\{Y_4, \ldots, Y_8\}$) and three dependent variables Y_9, Y_{10}, and Y_{11}. Log-linear models describe the associations among the independent variables, and logit models described the relations between the independent and dependent variables. The model contains several two-

way and three-way interactions. The appendix gives the details of the population model. Multiplying the population probability for each of the 2^{11} = 2048 response patterns by 1,000 produced the frequencies for all response patterns, amounting to a sample (of size $N = 1{,}000$) that is exactly in accordance with the population.

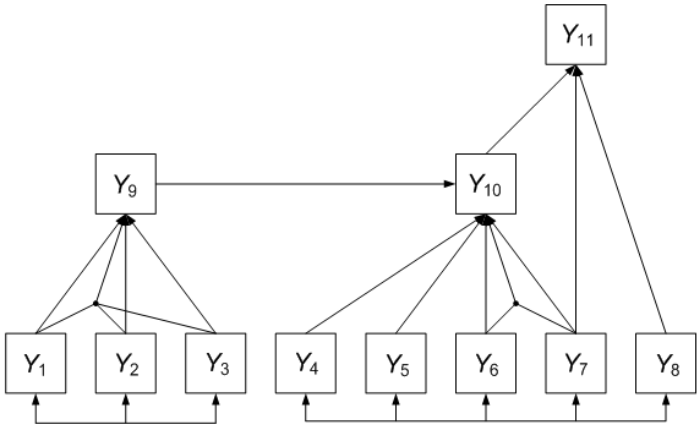

Figure 3.3: Population Model

We compared the true and the estimated marginal probabilities by a DLC model for three combinations of variables: $\{Y_9, Y_{10}\}$, $\{Y_8, Y_{11}\}$ and $\{Y_6, Y_7, Y_{10}\}$. Variables Y_6, Y_7, and Y_{10} have a three-way association and it is important to determine whether a DLC model is able to correctly pick up this complex association. The estimated marginal probabilities can be computed from the estimated DLC parameters. For example, the probabilities of Y_6, Y_7, and Y_{10} can be obtained by $\hat{P}(Y_6, Y_7, Y_{10}) = \sum_{s=1}^{K} \hat{P}(X = s)\, \hat{P}(Y_6|X = s)\hat{P}(Y_7|X = s)\hat{P}(Y_{10}|X = s)$. As an outcome variable we reported Pearson's chi-squared statistic for the differences between the true and the estimated expected frequencies for a sample size of $N = 1{,}000$. The degrees of freedom are equal to 3 for the two-way interactions, 7 for the three-way interactions, and 2047 for the entire density.

We estimated a DLC model using various decision rules for model selection in the local problem.

1. *Decision rules based on model fit*. Using this decision rule, an LC was split if the resulting LC*(2) model showed better fit than the LC*(1) model according to one of seven criteria. Six of the seven criteria were a combination of a minimum increase in the log-likelihood (levels 'at least 1 point increase' and 'at least 3 points increase'), and a maximum value of the highest standardized bivariate residual (levels 'unrestrictive', 'stop if all bivariate residuals are less than 1', and 'stop if all bivariate residuals are less than 3)'. The last level is to

keep splitting LCs as long as AIC decreases, which amounts to a minimum improvement of the log-likelihood equal to the number of additional parameters. We refer to Vermunt et al. (2008) for the discussion of preferring AIC over other relative fit statistics, such as AIC3 and BIC, in density estimation.

2. *Decision rules based on sample size*. Using this decision rule, an LC was only considered for splitting if it contained a minimal number of respondents. The three levels of minimal sample size were 0, 30, and 60.

The DLC model was estimated for all combinations of decision rules. The question whether the DLC model is able to accurately describe a complex density under ideal circumstances is investigated by examining the difference in true and estimated marginal probabilities under the least restrictive levels of the decision rules. This cell was used as an upper benchmark for investigating the effect of using more stringent levels of the decision rules. Note that levels of the decision rules of the upper benchmark should not be used in practice because one would also model all sampling fluctuations. Yet, comparing more stringent levels of decision rules to the upper benchmark is useful because it shows the relative decrease of precision in estimating the two-way and three-way interactions. Using a similar train of thought we used the independence model as a lower benchmark.

Results

For the upper bench mark (Table 3.2, first row), the values of the chi-squared statistics were very small compared to the degrees of freedom, indicating that the DLC can pick up complex associations in the data under ideal circumstances. Note that these values are too good to be true, and an additional simulation study (not tabulated) using data that were sampled from the model showed that the chi-square statistic—quantifying the difference between the estimated density and the population density—may actually increase if too many divisions are made (i.e., overfitting the data). As expected the lower benchmark (Table 3.2, last row) showed high values of the chi-squared statistics compared to the degrees of freedom, indicating that the independence model cannot describe complex associations in the data.

The more conservative decision rules especially deteriorated the model-fit for the two-way interaction of Y_8 and Y_{11}. The additional safeguards to have at least 30 respondents in each cell and to stop splitting when all standardized residuals are less than 1 did not affect the precision greatly. Other levels of the decision rules seriously deteriorated the density estimate, in particular choosing AIC as a criterion seems insufficient.

Table 3.2: *Chi-square Statistics for the Difference between the Estimated Frequencies using a DLC Model given Specific Decision Rules (Log-likelihood (L), Maximum Residual, and Minimum Class Size (N)) and the True Frequencies of Three Marginal Tables:* $y_9 y_{10}$, $y_8 y_{11}$, *and* $y_6 y_7 y_{10}$, *and the Total Data.*

Decision rules			Marginals			
LL	Residual	*N*	$y_9 y_{10}$	$y_8 y_{11}$	$y_6 y_7 y_{10}$	Total
1	0	0	.012	.065	.021	69.118
		30	.012	.897	.023	81.855
		60	.085	4.400	.208	187.064
1	1	0	.012	.053	.021	81.392
		30	.012	.896	.023	92.402
		60	.085	4.398	.208	192.870
1	5	0	.022	3.033	.470	164.787
		30	.022	3.033	.470	167.080
		60	.111	8.498	.828	219.414
5	0	0	.011	5.718	.059	174.113
		30	.011	5.718	.059	174.113
		60	.111	8.498	.828	219.414
5	1	0	.022	5.541	.477	180.048
		30	.022	5.541	.477	180.048
		60	.111	8.498	.828	219.414
5	5	0	.022	5.541	.477	180.048
		30	.022	5.541	.477	180.048
		60	.111	8.498	.828	219.414
AIC	0	0	.107	36.988	.857	337.374
		30	.010	37.639	.064	342.344
		60	.107	36.862	.858	331.970
Independence model			80.661	96.097	374.989	5740.297

3.4 Real-data Example

A possible application of a DLC model as a density estimation tool is multiple imputation (MI; Rubin, 1987). MI consists of creating *m* completed datasets by replacing the missing values in the data with plausible values *m* times. We analyzed a dataset from the ATLAS Cultural Tourism Research Project (Richards, 2010), a survey that addresses topics such as motivations, activities, and impressions of visitors of cultural sites and events. The dataset contained the scores of 4292 respondents on 79 categorical variables: 52 with 2 categories, 1 with 3, 19 with 5, 2 with 6, and 1 with 7, 8, 9, 10 and 17 categories, respectively. Complete information was available for only 794 respondents.

Table 3.3: *Variables Used in the Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 data.*

Variable		Categories	Number of Missing Values ($N = 4292$)
I want to find out more about the local culture	1	Totally disagree	154
	2	Disagree	
	3	Neutral	
	4	Agree	
	5	Totally agree	
Gender	1	Male	41
	2	Female	
Age	1	15 or younger	28
	2	16-19	
	3	20-29	
	4	30-39	
	5	40-49	
	6	50-59	
	7	60 or older	
Highest level of educational qualification	1	Primary school	62
	2	Secondary school	
	3	Vocational education	
	4	Bachelor's degree	
	5	Master's or doctoral degree	
Is your current occupation (or former)	1	Yes	149
	2	No	
Admission expenditure	1	0 - < 25 euro	2801
	2	25 - < 50 euro	
	3	50 - < 75 euro	
	4	75 - < 100 euro	
	5	≥ 100 euro	

The aim was to conduct an (adjacent-category) ordinal regression analysis (Agresti, 1990, pp. 286-288) using five explanatory variables to predict the responses to the question "I want to find out more about the local culture", which was answered on a five-point scale ranging from 1 (totally disagree) to 5 (totally agree). Table 3.3 provides detailed information on variables included in the regression. The missing data may cause bias in the parameter estimates and should be dealt with. We compared MI using the DLC model, MI using the standard LC model, and complete-case analysis as methods for solving the missing-data problem. If the DLC model performs well as a density estimation method, we expect the results of MI using a

DLC and LC model to be similar. For more information on multiple imputation and the comparison to complete-case analysis, see, for example, Schafer and Graham (2002).

By means of a simulation study, Vermunt et al. (2008) showed that MI using the LC model may yield approximately unbiased parameter estimates and standard errors. To achieve this, as many relevant variables as possible should be included in the LC model (Van der Palm et al., 2013a). Hence, preferably all 79 variables in the ATLAS dataset should be included. LC models can handle such a large numbers of variables, but it can be very time consuming (see Table 3.1, showing the computation time and model-fit of various LC models for the ATLAS dataset). With respect to bias, we expect the DLC model to perform similarly to the LC model, but much faster.

For both imputation methods (i.e., LC model and DLC model), we created 10 completed datasets using MI. First, we took 10 nonparametric bootstrap samples from the incomplete data. Second, we estimated an LC model and a DLC model for each bootstrap sample; the models were numbered 1 to 10. Third, for both imputation methods, completed dataset i $(i = 1, \dots, 10)$ was constructed by replacing the missing values of the original incomplete data by a value drawn from model i. This procedure takes the parameter uncertainty into account (Vermunt et al., 2008). For both methods, the regression analysis was conducted on all 10 completed datasets, and the 10 sets of parameter estimates were combined using Rubin's rules. The LC(65) and the DLC(95) model were used for MI. As a decision rule for splitting classes in the local problem we used a minimum increase of the log-likelihood of 1 point and a minimal sample size of 30 (consistent with the second DLC variant in the generated data example; row 2, Table 3.2).

We estimated two ordinal regression models to illustrate that MI should be preferred over complete-case analysis. In the first regression model Admission Expenditure was excluded as an explanatory variable, rendering 3950 complete cases, whereas in the second regression model Admission Expenditure was included, which reduced the number of complete cases to 1424. Table 3.4 shows the coefficients of the two ordinal regression models, estimated using complete-case analysis, and MI using the LC and the DLC model. For the first regression model (Table 3.4, upper panel), the parameter estimates and standard errors are rather similar across the methods, except for some differences in the parameter estimates for Education. Because there is only a small proportion of missing values in this analysis, it is not surprising that complete-case analysis and MI gave similar results. It is reassuring that for most regression coefficients, MI using the LC model and the DLC model provided similar estimates.

Table 3.4: *Parameter Estimates and Standard Errors of Two Ordinal Regression Analyses for the ATLAS Cultural Tourism Research Project 2003 Data using Complete-case Analysis, MI using an LC Model, and MI using a DLC Model (Minimum Increase of the Log-likelihood of 1 Point and a Minimal Sample Size of 30).*

			Multiple imputation			
	Complete-Case Analysis		MILC (*K*=65)		MIDLC (95)	
Predictor	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Gender	-.049	.026	-.052	.026	-.050	.025
Age	-.058	.010	-.062	.009	-.061	.009
Primary School	.000		.000		.000	
Secondary School	-.008	.098	-.039	.092	-.054	.093
Vocational Education	-.080	.098	-.098	.092	-.110	.094
Bachelor's Degree	-.067	.096	-.094	.089	-.105	.091
Master's or doctoral degree	-.091	.097	-.109	.091	-.124	.093
Occupation and culture	-.015	.030	-.017	.030	-.021	.029
Predictor						
Gender	-.077	.044	-.052	.026	-.050	.025
Age	-.082	.017	-.063	.009	-.061	.009
Primary School	.000		.000		.000	
Secondary School	-.110	.180	-.042	.092	-.058	.093
Vocational Education	-.152	.181	-.101	.092	-.114	.093
Bachelor's Degree	-.106	.176	-.097	.089	-.109	.091
Master's or doctoral degree	-.244	.179	-.113	.091	-.128	.093
Occupation and culture	-.041	.049	-.017	.030	-.021	.029
Admission Expenditure	.013	.014	.007	.012	.010	.012

For the second regression analysis (Table 3.4, lower panel), we found large differences in parameter estimates between complete-case analysis and MI. The estimates based on MI are similar in the two regressions, whereas the estimates based on complete-case analysis have changed: The estimated coefficients of age, gender, and education nearly doubled and all standard errors became larger. Although we cannot compare the estimates of the three methods to the population values, this result indicates that MI using the LC model and the DLC model performs well in this application. It is reassuring that the results based on the LC model and the DLC model are similar and largely concur with those of complete-case analysis

when the proportion of missingness is small and that the estimates are stable across the two regression analyses.

Table 3.5 shows the log-likelihood and the computation time for the LC model and the DLC model used for MI in the real-data example, plus for some alternative DLC models.

Table 3.5*: Data Log-likelihood yielded and Computation Time for MI using Four Different Models: The LC Model Consisted of 65 LCs, and the Three DLC Models used 62, 95, and 149 LCs, respectively.*

Method	Log-likelihood	Computation time
LC (65)	-216,043.09	8h12m
DLC(62)[1]	-217,990.25	0h47m
DLC(95)[2]	-213,337.94	1h02m
DLC(149)[3]	-205,340.37	1h06m

Note: 1 = minimum class-size at least $N = 60$, 2 = minimum class-size at least $N = 30$, 3 = minimum class-size at least $N = 10$.

The computation time for the standard LC model also includes the required computation time to estimate the LC models with fewer LCs. Table 3.5 shows that the DLC (95) (minimal improvement in the LL of 1 point and minimal sample size of 30) and DLC (149) (minimal improvement in the LL of 1 point and minimal sample size of 10) models yield a better fit than the LC (65) model, and in much less time.

3.5 Discussion

For density estimation the DLC model had three advantages over the standard LC model for density estimation. First, in the processes of finding a well fitting LC model, say LC(K), standard LC analysis requires estimating K models, whereas DLC analysis requires a single estimation procedure. Hence, it is no longer necessary to manually estimate and compare several models. Second, in standard LC analysis, the number of LCs is specified a priori, whereas in DLC analysis it is not; the number of LCs is increased during the estimation process until a sufficiently precise density estimate is obtained. Third, each LC model starts from scratch: the information in an LC(K) model is neglected when fitting an LC($K + 1$) model, whereas the DLC model is a sequence of small local problems and each local problem at Level $r + 1$ takes into account the information obtained at Level r. Due to this efficiency and relative simplicity, DLC estimation is much faster than LC estimation.

The generated data example showed that the DLC model is able to pick up two-way and three-way associations from a complex population model. The suggested decision rules for splitting classes worked well for the real-data example. However, additional research is required to further investigate the relationship between sample size, the specific decision rules, and precision of density estimation. Vermunt et al. (2008) found that over-fitting does not pose a big problem when using an LC model for density estimation. Therefore, the impact of over-fitting is expected to be limited for a DLC model as well. It should also be ascertained whether a minimum class size may be required to prevent over-fitting the data when dealing with sample fluctuation. In addition to over-fitting the data, it is also important to investigate the standard errors of the substantive model that may be estimated after MI, in relation to the specific decision rules in an extended simulation study.

The real-data example showed that a DLC model can easily be applied to a dataset with a large number of cases and polytomous variables. For a standard LC model with 65 LCs it took more than 8 hours to establish the best fitting model for this dataset, whereas a DLC model only required 1 hour and 2 minutes. In addition to being faster it yielded a better fit to the data. In a practical sense this makes a substantial difference for researchers that use an LC model as a density estimation tool. Our exemplary application underlines the benefits of a DLC model. If a researcher wants to use MI, the density of the data has to be estimated several times (10 times in this case). Hence, using a DLC model for MI instead of an LC model reduced the runtime for this dataset from 83h (10*8h18m) to 10h20m (10*1h2m).

DLC estimation has now been implemented in the software package Latent GOLD (Vermunt and Magidson, 2008) which makes it easier to apply the method. As an aside, we note that it is relatively easy to use multiple processing cores for the estimation of a DLC model because estimating the DLC model boils down to estimating a sequence of independent local problems. For the standard LC model, the processing load would have to be divided and delegated to each processor core within one estimation algorithm, which is more difficult and less efficient. For example, suppose a computer has four processor cores. After the first split (e.g., Figure 3.1), one processor cores can handle the estimation of the LCs beyond the first LC, and a second processing core can handle the estimation of the LCs beyond the second LC. After another split, the third processor core can be used. This makes the estimation process even faster.

Appendix

The densities used in Chapter 3 are described in terms of the realizations of Y, denoted by y. Let β_j denote a log-linear parameter value. The joint density of y_1, y_2, and y_3 is defined as,

$$P(y_{i1}, y_{i2}, y_{i3}) = \sum_{j=1}^{3} \beta_j y_{ij} + \sum_{j=1}^{2} \sum_{j'=j+1}^{3} \left(\beta_{jj'} y_{ij} y_{ij'} + \beta_{123} y_{i1} y_{i2} y_{i3}\right).$$

Hence, the joint density of y_1, y_2, and y_3 is in agreement with a saturated log-linear model containing all one-, two-, and three-variables associations. Table A1 shows the actual values of the parameters.

The joint density of $y_4, y_5, y_6, y_7,$ and y_8 is defined as

$$P(y_{i4}, y_{i5}, y_{i6}, y_{i7}, y_{i8}) = \sum_{j=1}^{5} \beta_j^4 y_{ij} + \sum_{j=1}^{4} \sum_{j'=j+1}^{5} \beta_{jj'}^5 y_{ij} y_{ij'},$$

and only contains two-way associations. Table A2 shows the actual values of the parameters.

Table A1: *Log-linear Parameters for the Density of* Y_1, Y_2, and Y_3.

Parameter	*Value*
β_1^1	*.2*
β_2^1	*-.6*
β_3^1	*.4*
β_4^1	*.2*
β_5^1	*.6*
β_{23}^2	*-.1*
β_{123}^3	*.2*

Table A2: *Log-linear Parameters for the Density of* Y_4, Y_5, Y_6, Y_7, and Y_8.

Parameter	*Value*	*Parameter*	*Value*	*Parameter*	*Value*
β_4^1	.2	β_{45}^2	.4	β_{57}^2	.6
β_5^1	-.6	β_{46}^2	-.2	β_{58}^2	-.2
β_6^1	.4	β_{47}^2	.6	β_{67}^2	.1
β_7^1	.2	β_{48}^2	-.3	β_{68}^2	-.2
β_8^1	.6	β_{56}^2	.8	β_{78}^2	.6

The conditional probabilities of the three dependent variables are defined to be in agreement with logit models, using effects coding for the parameters. Let β_j^q denote a logit regression parameter for the regression of dependent variable q on the jth independent variable. For dependent variable y_9,

$$\text{logit}(y_9) = \beta_0^{y_9} + \beta_1^{y_9} y_1 + \beta_2^{y_9} y_2 + \beta_3^{y_9} y_3 + \beta_{12}^{y_9} y_1 y_2 + \beta_{13}^{y_9} y_1 y_3 + \beta_{23}^{y_9} y_2 y_3 + \beta_{123}^{y_9} y_1 y_2 y_3,$$

for dependent variable y_{10},

$$\text{logit}(y_{10}) = \beta_0^{y_{10}} + \beta_9^{y_{10}} y_9 + \beta_4^{y_{10}} y_4 + \beta_5^{y_{10}} y_5 + \beta_6^{y_{10}} y_6 + \beta_7^{y_{10}} y_7 + \beta_8^{y_{10}} y_8 + \beta_{78}^{y_{10}} y_7 y_8,$$

and for dependent variable y_{11},

$$\text{logit}(y_{11}) = \beta_0^{y_{11}} + \beta_7^{y_{11}} y_7 + \beta_8^{y_{11}} y_8 + \beta_{10}^{y_{11}} y_{10}.$$

These relationships yield a complex density including three-way associations. Table A3 shows the values of the logistic regression parameters.

Table A3: *Logistic Regression Parameters for the Conditional Densities of* Y_9, Y_{10}, and Y_{11}.

Dependent Variable					
Y_9		Y_{10}		Y_{11}	
Parameter	*Value*	*Parameter*	*Value*	*Parameter*	*Value*
$\beta_0^{y_9}$	.0	$\beta_0^{y_{10}}$	.0	$\beta_0^{y_{11}}$	.0
$\beta_1^{y_9}$	.3	$\beta_9^{y_{10}}$	.5	$\beta_7^{y_{11}}$	-.6
$\beta_2^{y_9}$	.6	$\beta_4^{y_{10}}$	.6	$\beta_8^{y_{11}}$	.2
$\beta_3^{y_9}$	-.9	$\beta_5^{y_{10}}$	.1	$\beta_{10}^{y_{11}}$	.4
$\beta_{12}^{y_9}$	.3	$\beta_6^{y_{10}}$	-.4		
$\beta_{13}^{y_9}$	.5	$\beta_7^{y_{10}}$	-.8		
$\beta_{23}^{y_9}$	-.7	$\beta_{67}^{y_{10}}$	-.5		
$\beta_{123}^{y_9}$	-.2				

Chapter 4

Divisive Latent Class Modeling as an Incomplete-Data Method for Categorical Data

Abstract

We investigated the performance of the divisive latent class (DLC) model as an incomplete-data method. Relatively few incomplete-data methods are available for categorical data and the methods that are available suffer from serious practical issues. Maximum likelihood for incomplete data (MLID) and multiple imputation using a log-linear model (MILL) are the two most established methods for categorical data. Yet, MLID and MILL can handle only a few variables due to computational issues. Recently, multiple imputation using a latent class model (MILC) was introduced and was found to have a performance comparable to that of MLID and MILL in terms of bias and stability of parameter estimates. However, the required model-fit strategy for MILC may pose an obstacle to researchers and practitioners. Multiple imputation using a DLC model (MIDLC) solves the problems of MLID, MILL, and MILC: The method can handle a very large number of variables, is easier to use, and much faster to compute. However, the statistical properties of MIDLC have not been investigated yet. This article compares the performance of MIDLC with several commonly used incomplete-data methods using three studies. Results show that MIDLC generally has a performance comparable to that of MILC.

4.1 Introduction

Missing data is a commonly encountered issue within the social sciences that may prevent researchers from obtaining unbiased results in a statistical analysis. Missing-data must be dealt with in some way before the statistical model of interest can be estimated, for example, a linear or logistic regression model (from now on the statistical model of interest is referred to as the substantive model; the estimation of the substantive model is referred to as the substantive analysis). One way to deal with missing-data problems is to simply remove all the respondents from a dataset that have at least one missing value. Such a complete-case analysis, however, potentially produces biased parameter estimates in a substantive analysis (Little & Rubin, 2002; Schafer, 1997). Furthermore, because the observed part of each discarded incomplete response-vector is lost as well, the statistical power of the substantive analysis is unnecessarily reduced (Little & Rubin, 2002). Yet, ad hoc methods such as complete-case analysis remain the standard in the social sciences (Kim & Curry, 1977; Raghunathan, 2004; Klebanoff & Cole, 2008). This may partly be due to the fact that complete-case analysis is the default method in many statistical software programs. Furthermore, in comparison to the available incomplete-data methodology for continuous data, relatively few methods are available for categorical data (Van der Palm, Van der Ark, & Vermunt, 2013a).

Maximum likelihood for incomplete data (MLID; Dempster, Laird, & Rubin, 1977; Arbuckle, 1996; Allison, 2001; Little & Rubin, 2002; also known as full information maximum likelihood) is a very well known incomplete-data method for categorical data. MLID is known to produce unbiased results in substantive analyses, however, MLID may be computationally infeasible for datasets containing a large number of variables (Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008). MLID enables researchers to directly estimate a substantive model in the presence of missing data. A different approach to missing data is multiple imputation (MI; Rubin, 1987), and has become a widely established approach to deal with missing data problems. One important reason for this development is that MI is very practical (for an extensive discussion, see e.g., Schafer and Graham, 2002). MI can be described in four steps. First, one must define an imputation model. Second, using the estimated parameters of the imputation model, the missing values are filled in m times creating m completed datasets. Third, the substantive model is estimated on each of the m datasets. Finally, the results of the m substantive analyses are pooled using simple rules (Rubin, 1987) to obtain the final substantive results. MI is highly practical because the

handling of the missing data problem (steps 1 and 2) is separated from the substantive analysis (steps 3 and 4). The missing data problem only has to be addressed once (steps 1 and 2), and standard statistical techniques can then be used to estimate any substantive model. An important improvement of MI over single imputation is that the variation of the imputed values across the *m* completed datasets reflects uncertainty about the missing values (due to sampling error) as well as uncertainty about the parameters of the imputation model, which is a prerequisite for obtaining unbiased standard errors in the substantive analysis (Rubin, 1987). MI using a log-linear model (MILL; Schafer, 1997) is the gold standard of MI methods for categorical data and has been shown to produce unbiased parameter estimates in the substantive analysis. However, MILL can only handle a small number of variables (Schafer, 1997) which limits its practical use and, as we will discuss next, may indirectly cause biased results in the substantive analysis.

Advanced incomplete-data methods such as MLID and MILL are based on the assumption that the missing values are missing at random (MAR; Rubin, 1976; Little & Rubin, 2002). A violation of the MAR assumption has the potential to introduce bias in the results of a substantive analysis. If MLID is used to perform a substantive analysis, all available variables should be included in the substantive model because it may increase the degree to which the MAR assumption holds (Schafer, 1997). This requirement of including all available variables can be an obstacle to researchers because the computation time for each substantive analysis may be greatly increased and a smaller substantive model is typically preferred, containing only those variables that are relevant from a theoretical point of view. For MI methods, such as MILL, it is only necessary to include all available variables in the imputation model; subsequently, one can exclude any number of variables from the substantive model without affecting the MAR assumption (we elaborate upon MLID, MI, bias, and the MAR assumption in the incomplete-data section). However, to benefit from this advantage of MI over MLID, the imputation model must be able to include a potentially large number of variables, which is not the case with MILL.

Vermunt, Van Ginkel, Van der Ark, and Sijtsma (2008) introduced multiple imputation based on a latent class model (abbreviated as MILC). MILC can handle a very large number of variables and was found to have a performance comparable to that of MLID and MILL in terms of parameter bias (Vermunt et al., 2008; Gebregziabher & DeSantis, 2010; Van der Palm et al., 2013a). Thus, MILC allows researchers and practitioners to use a substantive model with a small number of variables without affecting the degree to which the MAR assumption holds. However, to use MILC, researchers have to find the best fitting latent

class model. A typical model-fit strategy is to estimate a 1-class model, a 2-class model, and so on, until the best fitting model has been found according to certain criteria. However, there are three issues with this model-fit strategy: (1) each subsequent latent class model will include an additional latent class, increasing the computation time; at some point the required computation time may become excessive, especially for datasets with a very large number of variables, (2) all of the latent class models have to be estimated and compared manually, which may be an obstacle for users of MILC and is relatively sensitive to human error, and (3) it has not yet been established how users of MILC should decide how many latent classes are enough.

Van der Palm, Van der Ark, and Vermunt (2013b) have developed a divisive latent class model that addresses the above three problems of MILC. A divisive latent class model consists of a sequence of 1- and 2-class models and each model builds on the results of the previous steps. Because a divisive latent class model is estimated sequentially the computation time is greatly reduced in comparison to a standard latent class model. As an example, Van der Palm et al. (2013b) analyzed a survey dataset with 79 variables; the computation time to find a sufficiently fitting latent class model was reduced from 8 hours and 12 minutes, to 1 hour and 2 minutes. In addition to faster results, a divisive latent class model produces the best fitting latent class model in a single run, without the need for human intervention during the estimation process; the optimal number of latent classes is automatically estimated within the procedure. Van der Palm et al. (2013b) showed that a divisive latent class model performs very well as a density estimator for categorical data. However, the divisive latent class model has not yet been investigated as a method for multiple imputation (abbreviated as MIDLC).

Vermunt et al. (2008) have already shown that a standard latent class model with a sufficient number of latent classes performs very well as an imputation model. In this paper we investigate the performance of a divisive latent class model as an incomplete-data method using three simulation studies. The goal is to determine under which settings a divisive latent class model yields a sufficiently fitting imputation model, resulting in unbiased results in the substantive analysis. Another incomplete-data method that can also handle a large number of variables is multivariate imputation using chained equations (MICE; Van Buuren & Groothuis-Oudshoorn, 2011), and we include this method for comparison to MILC and MIDLC. The first study concerns only a small number of dichotomous variables so that MLID and MILL could still be included for comparison to MIDLC. The second study concerns a larger number of trichotomous variables. The aim of Study 2 was to investigate

how MILC, MIDLC, complete-case analysis, and MICE perform for a larger number of polytomous variables. The third study concerns 11 variables with varying numbers of response categories; in order to increase the realism of the study the population model was defined using the association structure of an empirical dataset.

The remainder of this paper is structured as follows: first, we discuss the problem of incomplete-data and the incomplete-data methods; MILC and MIDLC are discussed in greater detail than the other methods to allow a thorough comparison of the two latent class approaches to the problem of missing data. Second, we present the results of three simulation studies. In the first study, for dichotomous data, we compare MLID, MILL, MILC, MIDLC, MICE, and complete-case analysis with respect to parameter bias, parameter stability, and bias of the standard errors. In the second study, for trichotomous data, we compare MILC, MIDLC, MICE, and complete-case analysis with respect to parameter bias, parameter stability, and bias of the standard errors. In the third study, we use the estimated associations within an empirical dataset as a population model, and compare MILC, MIDLC, MICE, and complete-case analysis with respect to parameter bias, parameter stability, and bias of the standard errors. Finally, we discuss the results of the three studies and give recommendations with regard to the use of the DLC model as an incomplete-data method.

4.2 Incomplete Data

Let $\mathbf{Y}$ denote the N x J matrix containing individuals indexed $i = 1, \dots, N$, with responses to variables indexed $j = 1, \dots, J$, and let $\boldsymbol{\theta}$ be the generic notation for the vector of unknown parameters of the joint distribution of $\mathbf{Y}$, denoted $P(\mathbf{Y}; \boldsymbol{\theta})$. Variable j is denoted by Y_j. Note that Y_j may either be a predictor variable or a dependent variable, depending on the substantive model. If confusion arises, we add the superscript p and d to indicate that a variable serves as a predictor variable or dependent variable, respectively. In case of missing data, the data matrix $\mathbf{Y}$ consists of an observed part, $\mathbf{Y}_{obs}$, and an unobserved part, $\mathbf{Y}_{mis}$, such that $\mathbf{Y} = (\boldsymbol{Y}_{obs}, \mathbf{Y}_{mis})$. Let $\mathbf{R}$ denote an indicator matrix with entries $r_{ij} = 0$ if the observation for person i on variable j is missing, and $r_{ij} = 1$ otherwise. It is crucial to consider how $\mathbf{R}$ relates to $\mathbf{Y}$—commonly referred to as the missingness mechanism—because the statistical properties of incomplete-data methods strongly depend on the characteristics of the mechanism (Little & Rubin, 2002). Advanced incomplete-data methods, including the ones discussed in this study, assume that the mechanism that governs the missing data is ignorable,

which means that two conditions should hold. First, the parameters of the missingness mechanism must be unrelated to the parameters of the substantive model (or, in case of MI, the parameters of the imputation model), which is a rather unrestrictive assumption (Schafer, 1997). Second, the MAR assumption must hold which states that the probability of a value being missing does not depend on the missing values themselves, or that such dependence can be fully explained by other (observed) variables in the dataset. Thus,

$$P(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}; \boldsymbol{\theta}) = P(\mathbf{R}|\mathbf{Y}_{\text{obs}}; \boldsymbol{\theta}). \quad (4.1)$$

One important issue in incomplete-data handling is the degree to which the MAR assumption holds because a violation may cause bias in the parameter estimates of the substantive model. In practice it is unlikely that the MAR assumption fully holds. However, as discussed by Schafer (1997), violations are typically minor and the extent to which the MAR assumption holds is a non-decreasing function of the number of variables that are included in the imputation model (in case of MI), or in the substantive model (in case of MLID). Thus, in general it is recommended to include all variables in every analysis because it potentially increases the degree to which the MAR assumption holds.

4.3 Incomplete Data Methods

4.3.1 Maximum Likelihood for Incomplete Data

MLID is a widely known method to obtain parameter estimates and standard errors of the substantive model directly in the presence of missing data (Little & Rubin, 2002). In contrast to complete-case analysis, MLID takes every single observed value into account. The substantive model can be an asymmetric model such as a logistic regression model, which describes the conditional distribution of the outcome variables given the predictor variables $P(\mathbf{Y}^{\text{d}}|\mathbf{Y}^{\text{p}}; \boldsymbol{\theta})$, or a symmetric model, such as a log-linear model or latent class model, which describe the joint distribution of all variables $P(\mathbf{Y}; \boldsymbol{\theta})$. To use MLID for categorical data, specialized software is usually required, such as LEM (Vermunt, 1997), or Mplus (Muthén & Muthén, 2010).

4.3.2 Multiple imputation

The incomplete-data methods that are based on MI use an imputation model to impute the incomplete data multiple times, and any substantive model can be estimated afterwards using the completed datasets. The MI methods differ in the imputation model that is used, and the way in which parameter uncertainty is taken into account during the imputation phase. The

imputation model describes the joint distribution of the data, $P(\mathbf{Y};\boldsymbol{\theta})$, and is used to draw imputation values to fill in the missing values. The general idea of MI is that the imputed values should behave neutrally in subsequent analyses, so that the associations in the data are not distorted. Typically, each missing value is imputed five times, creating five completed datasets. However, it is currently debated how many imputed datasets should be used (White, Royston, & Wood, 2010).

The primary goal of MI is to create imputation values that do not bias the results of subsequent substantive analyses. Thus, an imputation model should respect all associations whether existing only in the current sample or also at population level. Therefore, we argue that one does not have to be concerned about whether the imputation model over-fits the data. The substantive analysis is the appropriate moment to consider whether the estimated associations are systematic or merely sampling fluctuations.

One of the advantages of MI over MLID is that a substantive model can be estimated using the completed datasets that only includes a selection of the available variables, without affecting the degree to which the MAR assumption holds. MI facilitates such a practice because the incomplete-data problem has already been addressed in the imputation phase, completely utilizing $\mathbf{Y}_{\text{obs}}$ to impute $\mathbf{Y}_{\text{mis}}$, and creating the most favorable conditions for the MAR assumption to hold. However, to be able to make use of this advantage of MI, it must be possible to estimate the imputation model with a potentially large number of variables.

4.3.2.1 *Multiple imputation using a log-linear model*

MILL typically uses a saturated log-linear model as an imputation model. Consequently, when generating the imputation values all possible associations in the data are taken into account; for this reason MILL is the gold standard for MI of categorical data (Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008). However, MILL can only handle a small number of variables because of computational issues; the number of cells that have to be evaluated in a saturated log-linear model becomes far too large for datasets with more than a few variables. In such cases, only a selection of the available variables can be included in the imputation model. As a consequence, it is possible that the MAR assumption is affected because a number of variables may have been excluded that are part of the missingness mechanism, and Equation 4.1 may no longer hold. MILL can be applied using Latent GOLD (Vermunt & Magidson, 2008) or CAT (Schafer, 1997), which utilize a nonparametric bootstrapping and a Gibbs sampling approach, respectively, to account for parameter uncertainty.

4.3.2.2 *Multiple imputation using a latent class model*

MILC uses a standard LC model as an imputation model. Let X denote a discrete latent variable with K latent classes, index by k $(k = 1, \dots, K)$. Let $\boldsymbol{\pi}$ denote the vector of parameters of the latent class model; $\boldsymbol{\pi}$ can be divided into $\boldsymbol{\pi}_x$, the latent class proportions, and $\boldsymbol{\pi}_y$ the conditional response probabilities. Under a latent class model, joint distribution $P(\mathbf{y}_i; \boldsymbol{\pi})$ has the following form (Lazarsfeld, 1950; Goodman, 1974; Vermunt & Magidson, 2004):

$$P(\mathbf{y}_i;\ \boldsymbol{\pi}) = \sum_{k=1}^{K} P(X = k;\ \boldsymbol{\pi}_x) P\left(\mathbf{y}_i \middle| X = k;\ \boldsymbol{\pi}_y\right)$$

$$= \sum_{k=1}^{K} P(X = k;\ \boldsymbol{\pi}_x) \prod_{j=1}^{J} P\left(y_{ij} \middle| X = k;\ \boldsymbol{\pi}_{yj}\right).$$

In case of missing data only the observed part of $\mathbf{y}_i$ is used: $\mathbf{y}_{i,obs}$. However, for notational convenience we denote $\mathbf{y}_{i,obs}$ as $\mathbf{y}_i$. In the presence of missing data the LC model is defined as

$$P(\mathbf{y}_i;\ \boldsymbol{\pi}) = \sum_{k=1}^{K} P(X = k) \prod_{j=1}^{J} \left[P\left(y_{ij} \middle| X = k\right)\right]^{r_{ij}}.$$

If respondent *i* has a missing value for variable *j*, r_{ij} equals zero, and the conditional response probability is raised to the power zero for that item, setting it equal to 1. Technically, what occurs is that the missingness is summed out of the equation. Thus, if respondent *i* has a single missing value on variable *h*,

$$P(\mathbf{y}_i;\ \boldsymbol{\pi}) = \sum_{k=1}^{K} P(X = k) \sum_{c=1}^{C} P(y_{i1} | X = k) \cdot \dots \cdot P(y_{ih} = c | X = k) \cdot \dots \cdot P\left(y_{iJ} \middle| X = k\right)$$

$$= \sum_{k=1}^{K} P(X = k) \sum_{c=1}^{C} P(y_{i1} | X = k) \cdot \dots \cdot 1 \cdot \dots \cdot P\left(y_{iJ} \middle| X = k\right)$$

$$= \sum_{k=1}^{K} P(X = k) \prod_{j=1}^{J} \left[P\left(y_{ij} \middle| X = k\right)\right]^{r_{ij}}.$$

If the number of LCs is sufficient, the LC model is able to pick up the first, second, and higher order moments of the response variables, as is the case with all forms of mixture models (McLachlan & Peel, 2000). MILC can be applied using Latent GOLD (Vermunt &

Magidson, 2008), which utilizes a nonparametric bootstrap (Efron, Bradley, & Tibshirani, 1993) approach to account for parameter uncertainty.

The specific steps of MI using a standard LC model are as follows. First, obtain *M* nonparametric bootstrap samples. A nonparametric bootstrap sample is obtained by selecting *N* respondents from the original dataset with replacement, where *N* is the original sample size. Second, for each bootstrap sample estimate the LC model, yielding *M* sets of LC model parameters. Third, duplicate the original dataset *M* times, and for each copy use one of the *M* sets of parameters to impute the missing values. The imputation values for each of the *M* copies are obtained as follows:

For each respondent *i* (*i* in 1…*N*) that has at least one missing value,

(1) Calculate the probability of belonging to each of the *K* LCs (i.e., the posterior membership probabilities),

$$P(X = k|\,\mathbf{y}_i) = \frac{P(X = k)P(\mathbf{y}_i|X = k)}{\sum_{s=1}^{K} P(X = s)P(\mathbf{y}_i|X = s)} = \frac{P(X = k)\prod_{j=1}^{J}\left[P\left(y_{ij}\middle|X = k\right)\right]^{r_{ij}}}{\sum_{s=1}^{K} P(X = s)\prod_{j=1}^{J}\left[P\left(y_{ij}\middle|X = s\right)\right]^{r_{ij}}}.$$

(2) Randomly sample class membership using the posterior membership probabilities.

(3) Use the conditional response probabilities associated with the assigned LC to sample scores for the variables that have missing values. Because of the local independence assumption, sampling scores can be done separately for each variable.

One remaining question is how one should choose *K*, the number of LCs, when using MILC. Van der Palm et al. (2013a) suggested to let *K* be equal to the sum of the number of categories of all variables. The idea behind this decision rule is that for each additional variable included in the imputation model, more latent classes may be required to capture all associations. Furthermore, not only the number of variables is important but also the number of response categories of each manifest variable. For six dichotomous variables one should use 12 LCs, and for eleven trichotomous variables 33 LCs. We used this rule of thumb for Studies 1 and 2. For Study 3, however, this rule of thumb would result in an excessively large number of latent classes (due to the large number of response categories for several variables). Therefore, we used $K = 3 \cdot J$ as a decision rule for Study 3.

4.3.2.3 Multiple imputation using a divisive latent class model

We now discuss the details of performing MI using a DLC model (MIDLC). The DLC model estimation procedure constitutes a top-down clustering of respondents into LCs (Van der Palm et al. 2013b). It is obtained by estimating a series of one-class and two-class models. Thus, the whole sample starts in a single LC. The first question is whether the model-fit is improved by splitting the whole sample, creating the first two LCs. If so, the two new LCs are also checked to see whether further splits will improve the model fit. This process continues until sufficient model-fit has been achieved. The model-fit strategy for a standard LC model consists of estimating an LC model with *K* classes, *K*+1 classes, and so on, and each analysis starts from scratch. A DLC model, in contrast, increases the number of classes within one run until sufficient model fit has been obtained. Each one- and two-class model in the sequence builds on the results of the previous steps. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{ij}, \dots, y_{iJ})$ again be the observed part of the response vector of respondent *i* to the *J* manifest variables. Let X_z denote a latent variable at level *z*.

The estimation of a DLC model consists of the following iterative procedure:

(1) At start (Level 0, $z = 0$), the whole sample belongs to a single LC: $X_0 = 1$. Therefore, each individual belongs to LC: $X_0 = 1$ with a weight equal to 1, and $P(X_0 = 1|\, \mathbf{y}_i) = 1$.

(2) For class q at level z:

(2a) Estimate a one- and two-class model using $P(X_z = q|\, \mathbf{y}_i)$ as weights.

(2b) If the two-class model sufficiently improves the model-fit compared to the one-class model, class $X_z = q$ is split, creating classes $X_{z+1} = a$ and $X_{z+1} = b$ at level z+1. Otherwise, class $X_z = q$ is maintained, $X_{z+1} = X_z = q$, and the class is no longer considered for splitting in subsequent steps.

(2c) If class q was split, the probabilities for classes $X_{z+1} = a$ and $X_{z+1} = b$ are obtained by multiplying $P(X_z = q|\, \mathbf{y}_i)$ with $P(X_{z+1} = a|\, \mathbf{y}_i, X_z = q)$ and $P(X_{z+1} = b|\, \mathbf{y}_i, X_z = q)$, respectively. The latter two probabilities are simply the posterior probabilities obtained in the weighted latent class analysis for the sample associated with class $X_z = q$.

(3) Repeat step 2 for all classes at level *z*.

(4) Renumber the latent classes from 1 to Q (i.e., the total number of classes at level z), let $z = z + 1$, and repeat step 2 until all classes have either been split or maintained.

For technical details we refer to Van der Palm et al. (2013b).

A crucial part of the DLC model estimation procedure is the rule to decide whether the improvement in model fit—resulting from splitting an LC—is sufficient. Pilot studies were performed to address this issue in the context of missing data problems. Various decision rules have been discussed by Van der Palm et al. (2013b), however, because over-fit is not problematic in the context of missing data problems, we use a decision rule that only considers the improvement in the log-likelihood. Thus, the criterion is the difference in the log-likelihood of a 1- and 2-class model for a particular subsample (from now on referred to as DLL). Thus, if the DLL is large enough, according to the criterion one uses, an LC is split. For Study 1, the number of variables was rather small and we could use a DLL equal to 1 point, as suggested by Van der Palm et al. (2013b). However, in case of a larger number of variables and number of response categories, a DLL of 1 point yields far too many splits. Therefore, we performed several pilot studies to address this issue, and we found a rule of thumb that results in a sufficiently large number of LCs: let the required DLL be equal to $0.6 \cdot J$. We used this rule of thumb in Studies 2 and 3.

For the standard latent class model the nonparametric bootstrap is used to introduce variation in the imputation values to represent parameter uncertainty, so that unbiased standard errors are obtained for the parameter estimates in the substantive model. However, in a pilot study we found that MIDLC in combination with a nonparametric bootstrap actually introduces bias in the results of a subsequent substantive analysis. Furthermore, it was found that MIDLC yields unbiased standard errors without the nonparametric bootstrap. Therefore, we propose a slightly different procedure to perform MI using a DLC model, in comparison to a standard LC model.

The MI procedure for a DLC model can be summarized as follows,

(1) Estimate the DLC model for the original dataset.

(2) Compute the posterior membership probabilities using this single set of parameters.

(3) Duplicate the original dataset M times.

(4) Use the posterior membership probabilities to sample LC membership M times for each respondent.

(5) For each missing value that a respondent has, sample *M* responses using the conditional response probabilities of the *M* assigned LCs, and fill in the *M* copies of the original dataset.

4.3.2.4 *Multivariate imputation using chained equations*

MICE is a fully conditional specification (Van Buuren & Groothuis-Oudshoorn, 2011) method and, therefore, does not model the joint distribution of the variables directly. The imputation model is specified on a variable-by-variable basis using a separate conditional distribution for each incomplete variable. Effectively, MICE reduces the problem of finding one *J*-dimensional joint distribution to finding *J* univariate conditional distributions. Under certain conditions, a draw from each of the *J* conditional distributions is equivalent to a single draw from the joint distribution (Van Buuren, 2007), but it is not guaranteed. However, simulation studies suggest that the latter issue is unlikely to be problematic in practice (Van Buuren, Brand, Groothuis-Oudshoorn et al., 2006; Drechsler & Rassler, 2008). We used the R package `mice` (Van Buuren, 2007) to apply MICE, and used the default settings. Thus, logistic regression is used for dichotomous variables, and polytomous regression is used for variables with more than 2 categories.

4.3.2.5 *Complete-case analysis*

One of the best known ad-hoc methods is complete-case (CC) analysis, in which only the respondents without any missing values are used to estimate the substantive model. In contrast to the advanced incomplete-data methods such as MLID, MILL, MILC, and MICE, CC does not incorporate all available information. Therefore, the power of the substantive analysis is reduced, and the parameter estimates may be biased if the data are not missing completely at random (Little & Rubin, 2002; Schafer, 1997).

4.4 Study 1: Bias, Stability, and Bias of Standard Errors Produced by MLID, MILL, MILC, MIDLC, MICE, and Complete-Case Analysis for a Small Number of Dichotomous Variables.

In Study 1, we compared MLID, MILL, MILC, MIDLC (DLL = 1), MICE (logistic regression), and CC on the bias and stability of parameter estimates and bias of the standard errors. The number of dichotomous variables was kept small so that MLID and MILL could also be included for comparison to the other methods.

4.4.1 Method

4.4.1.1 *General Setup*

The simulation study was set up as follows. First, we sampled complete datasets from a population model. The population model was defined as follows for five dichotomous predictor variables $y_1, \ldots, y_5$, and one dichotomous outcome variable, y_6. The categories were coded 0 and 1 (dummy coding). Dummy coding was used because it is the most commonly used coding scheme for logistic regression models. The associations among the predictor variables were described by log-linear model (See Appendix A for the parameter values)

$$\log P(y_1, y_2, y_3, y_4, y_5) = \sum_{j=1}^{5} \beta_j y_j + \sum_{j=1}^{4} \sum_{k=j+1}^{5} \beta_{jk} y_j \, y_k. \quad (4.1)$$

Outcome variable y_6 was related to the predictor variables by logit model

$$\text{logit}(y_6) = \alpha - .8 \cdot y_1 + 1 \cdot y_2 + 1 \cdot y_3 + .4 \cdot y_4 + 1 \cdot y_5 - 1.20 \cdot y_2 y_3, \quad (4.2)$$

which contains main effects of the predictor variables as well as the interaction effect of y_2 and y_3. Complete datasets were created by sampling from $P(y_1, y_2, y_3, y_4, y_5)$ (Equation 4.1) and $P(y_6|y_1, y_2, y_3, y_4, y_5)$ (Equation 4.2).

Variables y_1 and y_2 had missing values that were MAR. Variable r_j indicated whether a score was missing for variable j, $r_j = 0$, or observed, $r_j = 1$. Missing values in y_1 were created using logistic regression model

$$\text{logit}(r_1) = \gamma_1 - 1 \cdot y_5 + 1 \cdot y_6 + .5 \cdot y_5 \, y_6, \quad (4.3)$$

and missing values in Y_2 were created using logistic regression model

$$\text{logit}(r_2) = \gamma_2 + .5 \cdot y_4 - .5 \cdot y_6 - 1 \, y_4 \, y_6. \quad (4.4)$$

The total percentage of missingness (one of the predictor variables in Study 1, to be discussed later) was manipulated by changing the intercepts (γ_1 and γ_2) of Equations 4.3 and 4.4, respectively. In this way, the percentage of missingness can be altered without changing the strength of associations between the predictor variables and the missingness indicator variables, r_1 and r_2.

For each incomplete dataset the six incomplete-data methods were used to estimate the following logistic regression model,

$$\text{logit}(y_6) = \alpha + \beta_1 \cdot y_1 + \beta_2 \cdot y_2 + \beta_3 \cdot y_3 + \beta_4 \cdot y_4 + .\beta_5 \cdot y_5 + \beta_{23} \cdot y_2 y_3. \quad (4.5)$$

Hence, the primary criterion for the incomplete data methods is how well the estimates of the parameters in Equation 4.5 approximate the population values defined in Equation 4.2. Note that the statistical model to be estimated is part of the population model as defined in Equations 4.1 and 4.2, and the missingness was defined to only depend on observed variables

included in both models. Thus, the MAR assumption holds by definition. The population model, as defined by Equations 4.1 and 4.2, contains a joint effect of y_2 and y_3 on y_6, implying that there is a three-variable association between $y_2, y_3,$ and y_6.

Three software packages were used for multiple imputation and parameter estimation. Data were generated using software package LEM (Vermunt, 1997), methods MILC, MIDLC and MILL were conducted using the software program Latent GOLD, and for MICE we used the R package MICE (Van Buuren, Groothuis-Oudshoorn, 2011). After MI, the substantive model, defined in Equation 4.5, was estimated using Latent GOLD. MLID and CC were applied directly to estimate the substantive model on each simulated dataset. For all incomplete-data methods we report the parameter estimates and standard errors of the substantive model.

4.4.1.2 *Experimental design*

To illustrate that using the nonparametric bootstrap in combination with MIDLC (abbreviated as MIDLC*) results in larger bias, we included this variant as well in the results. Thus, incomplete-data method was a within factor with seven levels: MLID, MILL, MILC, MIDLC, MIDLC*, MICE, and CC.

Percentage of missingness was a between factor with three levels: moderate (15% missingness), high (25% missingness), and extreme (40% missingness). The percentage of missingness was manipulated by varying parameters γ_1 and γ_2. For 15% missingness, $\gamma_1 = -2.03$ and $\gamma_2 = -1.59$, for 25% missingness, $\gamma_1 = -1.35$ and $\gamma_2 = -.92$, and for 40% missingness, $\gamma_1 = -.60$ and $\gamma_2 = -.19$.

Sample size was fixed to limit the scope of the first Study, and $N = 1000$. The seven incomplete-data methods were crossed with the three missingness percentages, yielding a 7×3 design, with 1000 replications for each level of missingness.

4.4.1.3 *Outcome variables*

The outcome variables were bias of parameter estimates, stability (the standard deviation of parameter estimates), and bias of the standard errors (Neyman & Pearson, 1933; Schafer & Graham, 2002). Let $\hat{\beta}_{bj}$ denote a parameter estimate of the jth variable (Equation 4.5) in replication b ($b = 1, \ldots, 1000$); we quantified bias as the relative percentage of bias over 1000 replications

$$\text{bias} = 100 \cdot \left(\frac{1}{1000} \sum_{b=1}^{1000} \left(\hat{\beta}_{bj} - \beta_j \right) \right) / \beta_j .$$

Stability, denoted by $sd(\hat{\beta}_j)$, was measured by the standard deviation of a parameter estimate across replications and was computed as

$$\text{sd}(\hat{\beta}_j) = \sqrt{\frac{1}{999} \sum_{b=1}^{1000} \left(\hat{\beta}_{bj} - \bar{\beta}_j \right)^2},$$

where $\bar{\beta}_j$ is the mean of $\hat{\beta}_{bj}$. Let $\text{se}(\hat{\beta}_{bj})$ denote the estimated standard error for parameter β_j in replication b. Bias of the standard errors was quantified as the percentage of relative bias; thus, we computed the mean difference between the estimated standard errors and the standard deviation of the parameter across replications, and divided it by the standard deviation to obtain relative SE bias

$$\text{SE bias} = 100 \cdot \left(\frac{1}{1000} \sum_{b=1}^{1000} \left(\text{se}(\hat{\beta}_{bj}) - \text{sd}(\hat{\beta}_j) \right) \right) / \text{sd}(\hat{\beta}_j) .$$

4.4.2 Results

4.4.2.1 *Bias*

Table 4.1 shows the results for bias for 15%, 25%, and 40% missingness. For 15% missingness, methods MLID, MILL, MILC, and MIDLC performed very well, generally producing small bias values. MIDLC*, MICE and CC produced rather large bias values for some of the coefficients. For 25% missingness, MLID, MILL, and MILC perform very well. MIDLC, produced small bias for all estimated coefficients; only for β_{23} MIDLC yielded a relative bias of 6.0%. In addition, because of the dummy coding, the estimates of coefficients β_2 and β_3 were affected as well, yielding larger bias. Methods MICE and CC produced rather large bias for several coefficients. For 40% missingness, the performance of MIDLC further deteriorated and the bias values for MICE and CC nearly doubled. MLID, MILL, and MILC performed very well. In general, MIDLC* produced larger bias than MIDLC.

Table 4.1: *Percentage of Bias in the Estimates of Six Logistic Regression Coefficients for Seven Incomplete-data Methods, for a Sample Size of 1000, and Three Different Percentages of Missingness.*

		Incomplete-data method						
Missingness	RC	MLID	MILL	MILC	MIDLC*	MIDLC	MICE	CC
15%	$\beta_1 = -.80$	-.4	-.6	-.4	.6	1.1	1.0	-.1
	$\beta_2 = 1.00$	2.0	1.9	2.1	-3.8	-0.8	-8.6	1.9
	$\beta_3 = 1.00$	1.5	1.4	1.6	-2.2	-0.4	-8.2	1.7
	$\beta_4 = .40$	1.3	1.5	1.3	2.0	0.8	0.8	0.8
	$\beta_5 = 1.00$	1.5	1.5	1.5	.2	0.8	0.3	16.7
	$\beta_{23} = -1.20$	-1.5	-1.3	-1.6	-4.2	1.8	14.1	-1.1
25%	$\beta_1 = -.80$	-0.5	-0.9	-0.9	-3.0	2.6	1.0	-1.0
	$\beta_2 = 1.00$	1.1	1.4	2.0	-9.4	-4.5	-16	1.3
	$\beta_3 = 1.00$	0.4	0.6	0.9	-5.8	-3.3	-15.4	0.1
	$\beta_4 = .40$	0.5	1.0	0.8	-2.2	-1.0	-0.3	2.0
	$\beta_5 = 1.00$	0.8	1.0	1.1	-.06	-0.6	-0.7	26.6
	$\beta_{23} = -1.20$	-0.5	-0.8	-1.3	-9.9	6.0	25.0	-0.2
40%	$\beta_1 = -.80$	-1.3	-1.1	-1.4	-6.4	2.8	1.4	-1.6
	$\beta_2 = 1.00$	2.2	2.6	3.2	-19.9	-7.8	-22.4	2.7
	$\beta_3 = 1.00$	2.0	2.5	2.9	-12.1	-4.2	-17.3	1.7
	$\beta_4 = .40$	1.0	1.3	1.3	-.1	-0.3	1.5	0.5
	$\beta_5 = 1.00$	1.2	1.6	1.8	-3.4	-1.4	-2.4	99.3
	$\beta_{23} = -1.20$	-2.3	-3.4	-3.6	-21.4	7.3	27.3	2.5

4.4.2.2 *Stability*

Table 4.2 shows the results for stability of the parameter estimates in the conditions with 15%, 25%, 40% missingness. The main result is that the differences in stability across the incomplete-data methods that have small bias values, are rather small. CC yielded parameter estimates with a relatively small stability.

Table 4.2: *Stability in the Estimates of Six Logistic Regression Coefficients for Seven Incomplete-data Methods, for a Sample Size of 1000, and Three Percentages of Missingness.*

		Incomplete-data method						
Missingness	RC	MLID	MILL	MILC	MIDLC*	MIDLC	MICE	CC
15%	$\beta_1 = -.80$	.163	.168	.169	.162	.167	.163	.180
	$\beta_2 = 1.00$	.221	.226	.227	.243	.220	.203	.236
	$\beta_3 = 1.00$	.214	.218	.219	.211	.213	.196	.240
	$\beta_4 = .40$	.153	.153	.154	.144	.152	.152	.179
	$\beta_5 = 1.00$	.156	.157	.157	.152	.156	.155	.194
	$\beta_{23} = -1.20$	.304	.312	.312	.303	.300	.259	.334
25%	$\beta_1 = -.80$	.164	.172	.172	.176	.165	.168	.193
	$\beta_2 = 1.00$	.234	.247	.247	.266	.241	.215	.256
	$\beta_3 = 1.00$	.223	.230	.230	.222	.221	.196	.271
	$\beta_4 = .40$	.154	.156	.156	.145	.154	.153	.202
	$\beta_5 = 1.00$	.146	.147	.148	.147	.146	.145	.202
	$\beta_{23} = -1.20$	.312	.327	.326	.318	.306	.240	.375
40%	$\beta_1 = -.80$	.168	.183	.184	.212	.170	.169	.201
	$\beta_2 = 1.00$	.256	.309	.309	.350	.263	.225	.265
	$\beta_3 = 1.00$	.231	.258	.261	.241	.222	.196	.271
	$\beta_4 = .40$	.150	.152	.153	.150	.148	.147	.213
	$\beta_5 = 1.00$	.160	.164	.164	.156	.159	.157	.227
	$\beta_{23} = -1.20$	.339	.392	.393	.345	.320	.247	.395

4.4.2.2 Bias in the Standard Errors

Table 4.3 shows the results for bias in the standard errors in the conditions with 15%, 25%, and 40% missingness. The main result is that all seven incomplete-data methods have very small bias in the estimated standard errors, and the differences across the seven methods are also very small. There appear to be no large difference between MIDLC and MIDLC*.

Table 4.3: *Percentage of Bias in the Standard Errors of the Estimates of Six Logistic Regression Coefficients for Seven Incomplete-data Methods, a Sample Size of 1000, and Three Percentages of Missingness.*

		Incomplete-data method						
Missingness	RC	MLID	MILL	MILC	MIDLC*	MIDLC	MICE	CC
15%	$\beta_1 = -.80$	-3.1	-6.5	-6.5	-1.8	-5.4	-1.8	-1.1
	$\beta_2 = 1.00$	1.4	-0.9	-1.3	-1.6	1.4	10.8	1.3
	$\beta_3 = 1.00$	-0.9	-2.8	-2.7	3.7	-0.5	8.2	0.0
	$\beta_4 = .40$	-3.9	-4.6	-4.5	1.6	-3.9	-3.9	-1.1
	$\beta_5 = 1.00$	-5.1	-5.1	-5.1	-1.8	-5.1	-4.5	-2.6
	$\beta_{23} = -1.20$	-0.7	-3.5	-3.2	3.7	0.3	16.6	0.0
25%	$\beta_1 = -.80$	3.0	-4.1	-3.5	-1.7	0.0	1.8	0.5
	$\beta_2 = 1.00$	2.1	-6.1	-6.1	3.9	-3.3	12.1	0.4
	$\beta_3 = 1.00$	-1.3	-5.2	-5.7	5.1	-1.4	11.2	0.0
	$\beta_4 = .40$	-3.9	-5.8	-5.8	2.9	-4.5	-3.3	-0.5
	$\beta_5 = 1.00$	2.7	-2.0	1.4	3.7	2.1	3.4	0.0
	$\beta_{23} = -1.20$	3.2	-4.3	-3.7	8.5	2.6	31.3	0.3
40%	$\beta_1 = -.80$	-0.6	-10.9	-10.9	-1.5	-3.5	-0.6	0.5
	$\beta_2 = 1.00$	0.4	-23	-23.3	3.5	-8.7	13.3	-0.4
	$\beta_3 = 1.00$	-0.9	-14.7	-16.1	8.8	-1.4	12.8	-0.4
	$\beta_4 = .40$	-0.7	-3.3	-3.3	3.9	-0.7	0.0	-0.5
	$\beta_5 = 1.00$	-4.4	-8.5	-8.5	4.3	-6.3	-4.5	-0.4
	$\beta_{23} = -1.20$	-1.2	-19.4	-20.4	16.1	-0.9	29.6	0.3

4.5 Study 2: Bias, Stability, and Bias of Standard Errors Produced by MILC, MIDLC, MICE, and Complete-Case Analysis for a Larger Number of Trichotomous Variables.

In Study 2 we compare MILC (K=33), MIDLC (DLL $= 0.6 \cdot J = 6.6$), MICE, and CC in terms of bias and stability of parameter estimates, and bias of the standard errors. Because of the larger number of variables ($J = 11$) and trichotomous instead of dichotomous data it is no longer practically feasible to apply benchmarks MLID and MILL. The aim of Study 2 is to investigate whether MILC, MIDLC, MICE, and CC perform well for polytomous categorical data and a large number of possible response patterns. In Study 1 the total possible number of response patterns equaled $2^6 = 64$, whereas in Study 2 it equaled $3^{11} = 177{,}147$. As for Study 1, the associations in the population model were defined to be complex to test whether the incomplete-data methods can correctly pick them up.

4.5.1 Method

4.5.1.1 *General set-up*

In Study 2, the population model (depicted in Figure 4.1) from which the complete datasets were sampled contained eight trichotomous predictor variables $(y_1, \ldots, y_8)$ and three trichotomous outcome variables $(y_9,\ y_{10}, \text{and}\ y_{11})$. The categories were coded 1, 2, and 3. The associations among the 11 variables are described by a path model for categorical data (Goodman, 1973) containing one-, two-, and three-way associations (see Appendix A for the chosen parameter values).

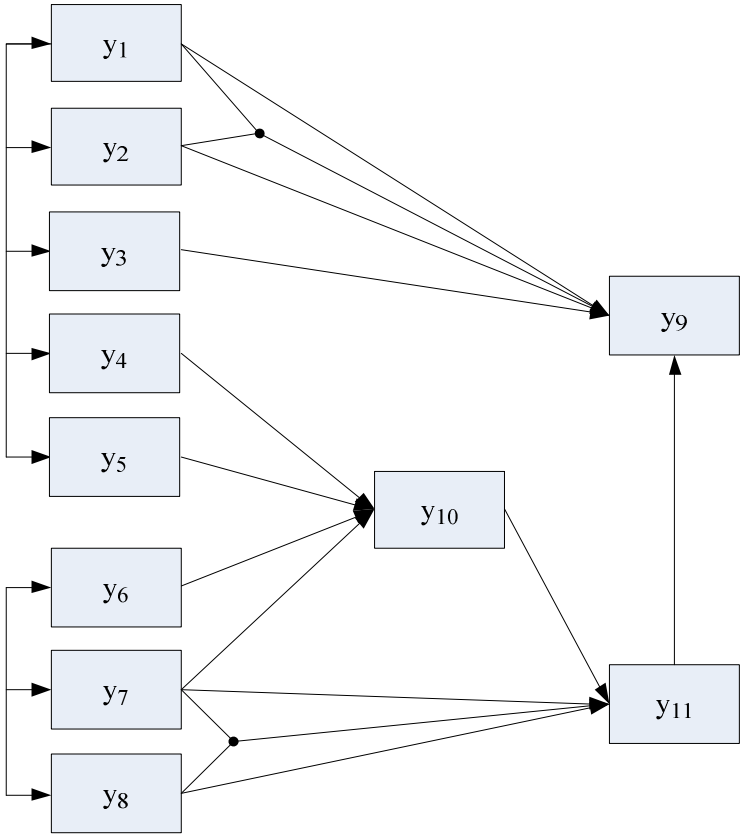

Figure 4.1: Population model of the second study. The model contains 11 trichotomous variables: y_1 through y_8 are predictor variables, and y_9 through y_{11} are outcome variables.

Variables $y_1, y_3, y_4, \text{and}\ y_{11}$ had missing values according to a MAR missingness mechanism; the other variables were completely observed. For y_1 and y_3 the missingness depended on y_2 and y_9. Let r indicate whether a value was observed $(r = 1)$ or not $(r = 0)$. Both for y_1 and y_3 the logit of r was $\text{logit}(r) = -5.06 - 2 \cdot y_9 + 3 \cdot y_2$, resulting in approximately 20% missing values for each of the two variables. Similarly, the missingness for y_4 and y_{11}

depended on y_7 and y_9. Both for y_4 and y_{11} the logit of r was $\text{logit}(r) = -5.50 + 3 \cdot y_9 - 1.5 \cdot y_7$, also resulting in approximately 20% missing values for both variables.

The substantive model was an adjacent category ordinal logit model (Agresti, 1990) for outcome variable Y_9, containing $Y_1, Y_3, Y_4,$ and Y_{11} as predictors. The logit equation has the form

$$\text{logit}(y_9 = j | y_9 = j - 1 \text{ or } y_9 = j) = \beta_{0j} + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \beta_4 y_4 + \beta_{12} y_1 y_2, \quad (4.6)$$

for $j = 2, 3$. Note that the substantive model is part of the population model (See Figure 4.1) and includes the main effects of the predictors of y_9, and the joint effect of $y_1$and y_2 on y_9.

Three software packages were used for data generation, incomplete-data handling, and estimating the substantive model. Complete and incomplete data were generated by LEM (Vermunt, 1997). The imputation phase of MILC, MIDLC, using Latent GOLD, and for MICE using the R package MICE. Latent GOLD was used to estimate the substantive model for complete-case analysis, and for MILC, MIDLC, and MICE, using the completed datasets.

4.5.1.2 *Experimental design*

For this study we varied sample size and incomplete-data method. Sample size had two levels: medium ($N = 500$) and large ($N = 1000$); incomplete-data method had four levels: MILC, MIDLC, MICE, and CC. This yields a 4 x 2 design. The outcome variables were equivalent to those in Study 1 (bias, stability, and bias of standard errors).

4.5.2 Results

4.5.2.1 *Bias*

Table 4.4 shows the bias for five coefficients of the population model for y_9, excluding the intercepts (Equation 4.6), for a sample size of 500 and 1000. We excluded the intercepts because any bias in the intercepts may also be due to bias in the other parameters and is, therefore, difficult to interpret. For $N = 500$, MIDLC had the smallest bias values followed by MILC, which produced remarkable bias values for parameters β_1, β_2, and β_{12}. For a sample size of 1000, MILC and MIDLC were the best performing methods, producing relatively small bias values. Methods MICE and CC produced very large bias in particular for the coefficients involved in the relatively complex interaction effect ($\beta_1, \beta_2,$ and β_{12}).

Table 4.4: *Percentage of Bias in the Estimates of Five Regression Coefficients for Four Incomplete-data Methods, and Two Sample Sizes (500, 1000).*

		Incomplete-data method			
Sample size	RC	MILC	MIDLC	MICE	CC
$N = 500$	$\beta_1 = -1.20$	14.3	8.1	52.4	23.6
	$\beta_2 = -1.35$	7.3	4.6	38.9	52.3
	$\beta_3 = .50$	0.6	5.2	10.2	4.6
	$\beta_{12} = .45$	-14.4	-8.2	-24.7	-23.3
	$\beta_5 = .35$	-9.4	-9.7	0.0	40.0
$N = 1000$	$\beta_1 = -1.20$	10.6	7.6	52.3	23.5
	$\beta_2 = -1.35$	5.4	4.3	38.7	52.1
	$\beta_3 = .50$	0.0	4.0	8.2	2.0
	$\beta_{12} = .45$	-10.7	-7.6	-24.2	-23.3
	$\beta_5 = .35$	-8.0	-9.1	-0.6	36.9

4.5.2.2 *Stability*

Table 4.5 shows the stability of the estimates produced by the four incomplete-data methods. MILC and MIDLC had a comparable stability. The estimates of MICE and CC that were strongly biased were also relatively stable.

Table 4.5: *Stability of the Estimates of Five Regression Coefficients for Four Incomplete-data Methods, and Two Sample Sizes (500, 1000).*

		Incomplete-data method			
Sample size	RC	MILC	MIDLC	MICE	CC
$N = 500$	$\beta_1 = -1.20$	.255	.263	.134	.377
	$\beta_2 = -1.35$	.240	.280	.134	.355
	$\beta_3 = .50$	.096	.101	.099	.135
	$\beta_{12} = .45$	.109	.112	.091	.164
	$\beta_5 = .35$	.091	.092	.095	.127
$N = 1000$	$\beta_1 = -1.20$	.190	.184	.092	.257
	$\beta_2 = -1.35$	.175	.174	.094	.249
	$\beta_3 = .50$	.070	.071	.068	.093
	$\beta_{12} = .45$	.082	.080	.063	.111
	$\beta_5 = .35$	.068	.066	.068	.090

4.5.2.3 *Bias of the standard errors*

Table 4.6 shows the bias in the standard errors for MILC, MIDLC, MICE, and CC. The four methods produced very small bias in the standard errors, and no substantial differences were observed across the methods.

Table 4.6: *Percentage of Bias in the Standard Errors of the Estimates of Five Regression Coefficients for Four Incomplete-data Methods, and Two Sample Sizes (500, 1000).*

		Incomplete-data method			
Sample size	RC	MILC	MIDLC	MICE	CC
$N = 500$	$\beta_1 = -1.20$	-6.7	-5.3	11.2	4.8
	$\beta_2 = -1.35$	-1.7	-2.1	8.2	3.1
	$\beta_3 = .50$	-10.4	-11.9	-9.1	-5.2
	$\beta_{12} = .45$	-2.8	-1.8	19.8	1.8
	$\beta_5 = .35$	-7.7	-5.4	-1.1	-3.9
$N = 1000$	$\beta_1 = -1.20$	-5.8	-4.3	15.2	7.8
	$\beta_2 = -1.35$	-2.3	-1.7	8.5	2.8
	$\beta_3 = .50$	-12.9	-11.3	-7.4	-3.2
	$\beta_{12} = .45$	-3.7	-2.5	22.2	4.5
	$\beta_5 = .35$	-7.4	-6.1	-5.9	-4.4

4.6 Study 3: Bias, Stability, and Bias of Standard Errors Produced by MLID, MILL, MILC, MIDLC, MICE, and Complete-Case Analysis for a Larger Number of Polytomous Items with a Population Model Based on an Empirical Dataset.

The aim of Study 3 is to investigate whether MILC, MIDLC, MICE, and CC perform well for polytomous categorical variables with varying numbers of categories and a population model that is based on the association structure observed in real data. Because the population model is defined by an empirical association structure, we argue that the realism of the simulated scenario is increased. We compared MILC ($K = J \cdot 3 = 33$), MIDLC (DLL $= 0.6 \cdot J = 6.6$), MICE, and CC in terms of bias and stability of parameter estimates, and bias of the estimated standard errors. Benchmarks MLID and MILL could no longer be applied because of the larger number of variables ($J = 11$). The empirical dataset that we analyzed originates from the Psychological Contracts across Employment Situation (PSYCONES) project (Dutch and Belgian sample; European Commission, 2006). The PSYCONES project administered a questionnaire to respondents containing statements and questions such as ‘I am happy with my current job’, ‘I have to work under pressure’, and ‘During the last 12 months, how often have you been absent from work due to health reasons?’.

4.6.1.1 *General set up*

The dataset contained scores of 1442 respondents on 203 variables, most of which were categorical. We selected 11 categorical variables to define the population model: ten predictor variables ($y_2, \ldots, y_{11}$), and one outcome variable y_1, as summarized in Table 4.7.

Table 4.7: *Variables Used in the Ordinal Regression Model for the PSYCONES Dataset.*

Variable	Description		Categories
Y_1	'I am happy with my work'	5	1=Completely disagree, …, 5=Completely agree.
Y_2	'Position at work'	6	1=Unskilled blue collar worker, …, 6=Management or director.
Y_3	'My job requires creativity'	5	1=Completely disagree, …, 5=Completely agree.
Y_4	'I have to work under pressure'	5	1=Seldom or never, …, 5=Very frequently or always.
Y_5	'I have to work late'	5	1=Seldom or never, …, 5=Very frequently or always.
Y_6	'During the last week, how well did you perform in terms of making errors?'	5	1=Very poorly, …, 5=Very well.
Y_7	'My supervisor helps me to get my job done'	5	1=Completely disagree, …, 5=Completely agree.
Y_8	'My supervisor cares about my opinion'	5	1=Completely disagree, …, 5=Completely agree.
Y_9	'How satisfied are you with your life at the moment?'	7	1=Very dissatisfied, …, 7=Very satisfied.
Y_{10}	'Absence during the last 12 months due to health problems'	5	1=Never, …, 5=More than five times.
Y_{11}	'Gender'	2	1=Female, 2=Male.

The dependent variable, 'I am happy with my current job', was predicted by variables such as 'My supervisor cares about my opinion', and 'I have to work under pressure'. To completely control the missingness mechanism in this study, we excluded all cases with at least one missing value, resulting in a dataset with 1218 complete cases. The aim was to estimate an adjacent category ordinal logit model (Agresti, 1990) to obtain the parameter estimates for the original dataset, and to use this set of parameters as a population model to generate data. The logit equation has the form

$$\text{logit}(y_1 = j | y_1 = j-1 \text{ or } y_1 = j) = \beta_{0j} + \beta_2 y_2 + \beta_3 y_3 + \beta_4 y_4 + \beta_5 y_5 + \beta_6 y_6 + \beta_7 y_7 + \beta_8 y_8 + \beta_9 y_9 + \beta_{10} y_{10} + \beta_{11} y_{11} + \beta_{2,9} y_2 y_9. \quad (4.7)$$

The "population parameter values" (i.e., the estimated parameter values we obtained before creating missingness) for the ordinal logit model are depicted in Table 4.8.

Table 4.8: *Population Parameter Values for the Ordinal Regression Model Estimated for the PSYCONES Dataset.*

Parameter	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	$\beta_{2,9}$
Value	-.411	.223	-.102	-.075	-.108	.108	.246	-.062	-.161	-.140	.082

Generated datasets were obtained by randomly selecting *N* respondents from the original dataset, with replacement, and with *N* equal to 1000. We defined a MAR missingness mechanism as follows to create missing values in each sample that was drawn from the original dataset. As for Studies 1 and 2, we defined two logistic regression models to introduce missingness for variables $y_4, y_7, y_9,$ and y_{10}. Variables r_j indicated whether a score was missing for variable $j, r_j = 0$, or observed, $r_j = 1$. Missing values in y_4 and y_7 were created using logistic regression model

$$\text{logit}(r_4) = \text{logit}(r_7) = -2.308 - .25 \cdot y_1 + .3 \cdot y_2 + .05 \cdot y_1 y_2,$$

and missing values in y_9 and y_{10} were created using logistic regression model

$$\text{logit}(r_9) = \text{logit}(r_{10}) = -3.79 + .25 \cdot y_1 + .6 \cdot y_8 - .05 \cdot y_1 y_8.$$

Each variable with missingness was defined to have 20% missingness on average.

For each generated dataset, MILC, MIDLC, MICE, and CC were used to estimate the regression model as defined in Equation 4.7. Two software packages were used for multiple imputation and parameter estimation. Data were generated using R (R Development Core Team, 2013), methods MILC and MILL were conducted using the software program Latent GOLD, MIDLC was applied using R and Latent GOLD, and for MICE we used the R package MICE (Van Buuren, Groothuis-Oudshoorn, 2011). After multiple imputation, the substantive model (Equation 4.7) was estimated using Latent GOLD.

4.6.1.2 *Design*

For this study we only varied incomplete-data method and used a fixed sample size of 1000. Incomplete-data method had four levels: MILC, MIDLC, MICE, and CC. The outcome

variables were equivalent to those in Studies 1 and 2 (bias, stability, and bias of standard errors).

4.6.2 Results

4.6.2.1 *Bias*

Table 4.9 shows the bias for 11 coefficients of the population model, excluding the intercepts (Equation 4.7), for a sample size of 1000. MILC and MIDLC produced relatively large bias in the estimate of β_2, and MIDLC for β_8 as well. Otherwise, MILC and MIDLC performed very well yielding small bias values. MICE and CC produced large bias in several parameter estimates.

Table 4.9: *Percentage of Bias in the Parameter Estimates of Eleven Regression Coefficients for Four Incomplete-data Methods.*

		Incomplete-data method			
Sample size	RC	MILC	MIDLC	MICE	CC
N=1000	$\beta_2 = -.411$	16.1	23.6	21.2	-6.3
	$\beta_3 = .228$	-3.1	-3.5	-3.1	3.1
	$\beta_4 = -.102$	4.9	13.7	2.9	-16.7
	$\beta_5 = -.075$	-4.0	-20.0	0.0	-9.3
	$\beta_6 = .108$	13.0	24.1	14.8	26.9
	$\beta_7 = .108$	7.4	-20.4	2.8	9.3
	$\beta_8 = .246$	2.0	11.4	2.8	10.2
	$\beta_9 = -.062$	41.9	46.8	61.3	4.8
	$\beta_{10} = -.161$	-1.2	4.3	-5.0	-3.1
	$\beta_{11} = -.140$	0.7	1.4	-1.4	-7.9
	$\beta_{2,9} = .083$	-15.7	-24.1	-19.3	3.6

4.6.2.2 *Stability*

Table 4.10 shows the stability of the estimates produced by the four incomplete-data methods. MIDLC generally produced the most stable parameter estimates. MILC and MICE yielded a stability of parameter estimates comparable to MIDLC. CC produced very unstable parameter estimates.

Table 4.10: *Stability of the Parameter Estimates of Eleven Regression Coefficients for Four Incomplete-data Methods.*

		Incomplete-data method			
Sample size	RC	MILC	MIDLC	MICE	CC
N=1000	$\beta_2 = -.411$	.147	.141	.138	.246
	$\beta_3 = .228$	.048	.047	.049	.076
	$\beta_4 = -.102$	.045	.040	.050	.068
	$\beta_5 = -.075$	.046	.045	.047	.073
	$\beta_6 = .108$	.076	.075	.078	.117
	$\beta_7 = .108$	.044	.037	.050	.067
	$\beta_8 = .246$	.049	.048	.052	.079
	$\beta_9 = -.062$	.092	.090	.089	.143
	$\beta_{10} = -.161$	.044	.041	.048	.069
	$\beta_{11} = -.140$	.082	.080	.083	.132
	$\beta_{2,9} = .083$	.027	.026	.025	.045

4.6.2.3 *Bias of the standard errors*

Table 4.11 depicts the bias in the standard errors for MILC, MIDLC, MICE, and CC. In general, MILC, MIDLC, and MICE produced small bias in estimating the standard errors for all parameters. MILC and MICE slightly underestimated the standard errors for most parameter estimates, whereas MIDLC produced a more equal number of slight over- and underestimates of the standard errors. CC yielded relatively large underestimates of the standard errors.

Table 4.11: *Percentage of Bias in the Standard Errors for Eleven Regression Coefficients for Four Incomplete-Data Methods.*

		Incomplete-data method			
Sample size	RC	MILC	MIDLC	MICE	CC
N=1000	$\beta_2 = -.411$	-2.7	12.1	4.3	-13.8
	$\beta_3 = .228$	-10.4	-10.6	-10.2	-14.5
	$\beta_4 = -.102$	6.7	20.0	-2.0	-2.9
	$\beta_5 = -.075$	-4.3	-6.7	-6.4	-8.2
	$\beta_6 = .108$	-7.9	-9.3	-9.0	-8.5
	$\beta_7 = .108$	-2.3	18.9	-12.0	-7.5
	$\beta_8 = .246$	-6.1	-6.3	-7.7	-11.4
	$\beta_9 = -.062$	-6.5	-2.2	-2.2	-15.4
	$\beta_{10} = -.161$	2.3	12.2	4.2	-5.8
	$\beta_{11} = -.140$	-2.4	-2.5	-3.6	5.3
	$\beta_{2,9} = .083$	-3.7	11.5	4.0	-15.6

4.7 Discussion

In this paper we performed an initial investigation of the DLC model as an incomplete-data method. Results of Study 1 showed that MIDLC generally had a performance equal to MILC, although MIDLC yielded a slightly biased estimate of the complex interaction effect. In Study 2 we found that MIDLC outperformed MILC, generally yielding smaller bias in the parameter estimates. In Study 3, differences between MILC and MIDLC were small in terms of parameter bias.

In the first study MILC and MIDLC were also compared to gold standards MLID and MILL, and we found that the four methods generally have a comparable performance; it should be noted that the performance of MIDLC deteriorated somewhat as a function of the percentage of missing data. In Studies 1, 2, and 3, MILC and MIDLC were compared to MICE and complete case analysis. From the results of Studies 1 and 2, it can be concluded that in comparison to MILC and MIDLC, MICE and complete case analysis produced large bias in several parameter estimates. In Study 3 the differences between these four methods were smaller. Based on the results of the three studies we conclude that MIDLC can be preferred over the default implementation of MICE and complete case analysis, as the latter two potentially yield large bias in the parameter estimates.

Additional research is required to investigate why the performance of MIDLC is different for dichotomous and trichotomous data, and why the DLC model does not properly capture the complex interaction term in case of dichotomous data. Thus, with respect to MIDLC, it is clear that there is room for improvement, especially for dichotomous data. A more extensive simulation study would be highly useful to assess whether better decision rules can be formulated, and how they affect the performance of MIDLC.

Another issue with MIDLC is whether it should be combined with a nonparametric bootstrap. As shown in the three studies, MIDLC does not appear to require the nonparametric bootstrap to obtain unbiased standard errors. A possible explanation for this is that a DLC model typically contains a (much) larger number of LCs than a standard LC model to obtain the same level of precision (i.e., the extent to which all associations are captured). This may already introduce sufficient additional variation across the multiple imputations, resulting in unbiased standard errors when using a DLC model for MI. However, further research is needed to clarify this issue.

Chapter 5

Test-score Reliability for Multidimensional Educational Tests

Abstract

Most items in an educational test require proficiency on multiple abilities, skills, and knowledge domains and, therefore, when administered to students yield multidimensional data. Reliability estimation methods for multidimensional data are available but suffer from several practical problems. We propose the adapted latent class reliability coefficient that solves these problems and is particularly suited for multidimensional data. Results showed that the adapted latent class reliability coefficient produces a less biased reliability estimate than other methods in a wide range of scenarios involving multidimensional data.

This chapter has been submitted for publication.

Introduction

In this study, we extend a test-score reliability estimation method for multidimensional educational tests. The extended method is based on a method recently proposed by Van der Ark, Van der Palm, and Sijtsma (2011). Both methods are based on the latent class model (Lazarsfeld, 1950; also see, e.g., McCutcheon, 1987; Goodman, 1974; Hagenaars & McCutcheon, 2002). Many methods exist to estimate test-score reliability but most have not been designed for the case that the data are multidimensional and thus may provide estimates that are biased relative to the reliability. Hence, methods adapted to multidimensionality may be more appropriate, in particular when test scores based on the multidimensional data are used to make important decisions about individuals.

Reise, Waller and Comrey (2000, p. 294) argued that, unless a test measures a very narrow construct such as the skill of two-digit number addition (e.g., 27 + 36 = ?), it is unlikely that the test is unidimensional. In educational testing, most tests are multidimensional because they typically require knowledge of multiple topics or multiple sub-abilities or skills. For example, an educational test on the subject of test theory typically contains items that ask for knowledge about different topics such as history of testing, item writing, reliability, validity, and ethics of testing. Another possibility is that each item simultaneously requires arithmetic ability, knowledge of elementary statistics, and knowledge of classical test theory. Test users usually assign one grade to each student, irrespective of data multidimensionality. Hence, it is important to know the reliability of the test scores. For multidimensional test data, many reliability estimation methods produce gross underestimates of test-score reliability (Komaroff, 1997; Murphy & DeShon, 2000; Osburn, 2000; Raykov, 1998; 2001; Van der Ark et al., 2011; Zimmerman, Zumbo, & Lalonde, 1993).

Methods to estimate test-score reliability tailored to multidimensional data are stratified alpha (Cronbach, Schoneman, & McKie, 1965) and maximal reliability (Li, Rosenthal, & Rubin, 1996). Previous research showed that stratified alpha and maximal reliability estimated test-score reliability well for multidimensional data (Osburn, 2000; Kamata, Turhan, & Darandari, 2003). To use these methods, the researcher must divide the item set into subsets that each assesses a different knowledge domain, sub-ability or skill. The item division may be problematic when one does not know which knowledge domains, sub-abilities or skills underlie test performance. Even more problematic is the situation in which different items are driven by different subsets of knowledge domains, sub-abilities or skills, so that an item is a member of different item subsets and a division based on items measuring

one knowledge domain, sub-ability or skill is impossible. If an item is assigned to the wrong subset, stratified alpha and maximal reliability underestimate reliability (Kamata et al., 2003). This study proposes a method that circumvents these problems and produces almost unbiased test-score reliability estimates when test data are multidimensional. This is a variation of the latent class reliability coefficient (LCRC; Van der Ark et al., 2011).

LCRC yields practically unbiased reliability estimates (Van der Ark et al., 2011). LCRC performed well for dichotomous items and polytomous items, small and large samples, items having equal and unequal discrimination parameters, and multidimensional data. LCRC does not require the assignment of items to item subsets. Hence, LCRC is a promising candidate for estimating test-score reliability for multidimensional educational test data. However, method LCRC has two practical shortcomings: (1) LCRC may require excessive computation time for tests containing more than 30 items; and (2) to compute LCRC, the user must manually compare the fit of each model in a series of latent class models. Obviously, in this respect the method is user-unfriendly and increases the risk of error.

We propose an improved version of method LCRC, denoted LCRC*, which solves the problems of long computation time for long tests and user-unfriendliness due to having to evaluate intermediate latent class results. For LCRC, Van der Ark et al. (2011) used a traditional latent class model to estimate the joint distribution of the item scores. For LCRC*, we propose to use a divisive latent class (DLC; Van der Palm, Van der Ark, & Vermunt, 2013b) model. The DLC model greatly reduces computation time and requires only a single run to obtain the best fitting latent class model, thus relieving the researcher from having to make difficult and often arbitrary decisions.

This study investigated which specific model-fit strategy should be used for LCRC*, and compared LCRC* to LCRC. Also, LCRC and LCRC* were compared to the following lower bounds to the reliability: coefficient alpha (e.g., Cronbach, 1951), coefficient lambda2 (Guttman, 1945), and the greatest lower bound (GLB; Bentler & Woodward, 1980; Ten Berge, Snijders, & Zegers, 1981; Woodhouse & Jackson, 1977).

This article is organized as follows. First, we discuss the definition of reliability in the classical test theory framework. Second, we discuss LCRC, LCRC*, alpha, lambda2, and GLB. Third, we discuss a simulation study that compares LCRC, LCRC*, alpha, lambda2, and GLB with respect to bias and accuracy relative to the test-score reliability. Fourth, we discuss the implications for reliability estimation.

Reliability Theory

Let a test contain J items and let test score X denote the sum of the J item scores; that is, $X = \sum_{j=1}^{J} X_j$. The propensity distribution of an individual is his distribution of test scores across hypothetical, independent repetitions of the test (Lord & Novick, 1968, pp. 29-30). True score T is the individual's expected test score over independent repetitions; hence, it is the mean of his propensity distribution. The deviation of a person's test score from his true score is the random measurement error, E. The classical test model equals $E = X - T$. Measurement error correlates zero with any other variable Y in which it is not included, so that, using ρ to denote the product-moment correlation, $\rho_{EY} = 0$. Let σ^2 denote the variance of a variable, it can be shown that from the assumptions of classical test theory it follows that $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$.

Two tests X and X' are parallel if (1) for each person the true scores on the two tests are equal, $T = T'$, and (2) the tests have equal variance, $\sigma_X^2 = \sigma_{X'}^2$ (Lord & Novick, 1968, p. 48). Parallel tests can be considered a formalization of independent repetitions of a test. Test-score reliability is defined as the product-moment correlation between two parallel tests, and can be shown to equal the proportion of true-score variance on each of the tests,

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_{T'}^2}{\sigma_{X'}^2}. \tag{5.1}$$

In practice, two parallel tests usually are unavailable and true-score variance is unobservable. Hence, under these conditions reliability as defined in Equation (5.1) cannot be estimated from real data. Instead, several reliability estimation methods have been proposed that only use one set of data to approximate the correlation between two parallel forms (Equation 5.1).

We use the statistical framework for test-score reliability that Van der Ark et al. (2011) proposed. Let the sample size be denoted N and assume that n_j persons answered item j correctly and n_{ij} persons answered both items i and j correctly. Let π_j denote the probability that a randomly drawn person answers item j ($j = 1, \ldots, J$) correctly, and let $p_j = n_j/N$ be its sample estimate. Let π_{ij} denote the probability that a randomly drawn person answers both items i and j correctly, and let $p_{ij} = n_{ij}/N$ ($i \neq j$) be its sample estimate. For $i = j$, π_{ii} is the probability that a randomly drawn person answers item i correctly in two independent repetitions. But, as each item has been administered only once, this probability is

unobservable and has to be estimated, for example, using procedures proposed by Sijtsma and Molenaar (1987).

For dichotomous items, Molenaar and Sijtsma (1988) showed that reliability (Equation 5.1) can be written as,

$$\rho_{XX'} = \frac{\sum_{i=1}^{J}\sum_{j=1}^{J}[\pi_{ij} - \pi_i\pi_j]}{\sigma_X^2}. \quad (5.2)$$

The ratio in Equation 5.2 can be divided into the sum of two ratios,

$$\rho_{XX'} = \frac{\sum\sum_{i\neq j}[\pi_{ij} - \pi_i\pi_j]}{\sigma_X^2} + \frac{\sum_i[\pi_{ii} - \pi_i\pi_i]}{\sigma_X^2}. \quad (5.3)$$

The numerator of the first ratio involves observable quantities but the numerator of the second ratio contains the unobservable joint probability π_{ii}. Methods LCRC and LCRC* are discussed using the definition of reliability in Equation 5.3. The methods provide a solution to estimate unobservable probability π_{ii} from one dataset.

Method LCRC

The latent class model is the basis of method LCRC. Let ξ denote the discrete latent variable and assume ξ has K classes. Conditional on ξ, the manifest variables are statistically independent; this is also known as conditional or local independence. The latent class model describes the joint probability distribution of the J item scores as,

$$P(X_1 = x_1, \ldots, X_J = x_J) = \sum_{k=1}^{K} P(\xi = k)\prod_{j=1}^{J} P(X_j = x_j|\xi = k). \quad (5.4)$$

The parameters of a latent class model are the marginal class probabilities, $P(\xi = k)$, and the conditional response probabilities, $P(X_j = x_j|\xi = k)$. Under a latent class model with K latent classes (Equation 5.4), the unobservable probability, π_{ii}, in Equation 5.3 equals

$$\pi_{ii} \equiv P(X_i = 1, X_i = 1) = \sum_{k=1}^{K} P(\xi = k)[P(X_i = 1|\xi = k)]^2. \quad (5.5)$$

Method LCRC replaces π_{ii} in Equation 5.3 by the right-hand side of Equation 5.5. Hence, method LCRC equals test-score reliability $\rho_{XX'}$ only if the latent class model with K latent classes perfectly describes $P(X_1 = x_1, \ldots, X_J = x_J)$.

In samples, π_{ii} is estimated using a latent class model that fits the data. Van der Ark et al. (2011) used the following strategy to obtain the latent class model. Relative fit index AIC3 (Bozdogan, 1987) is used to compare the fit of different latent class models: A lower value of

AIC3 corresponds to a better fitting latent class model. A series of latent class models are estimated, starting with one latent class and increasing the number of latent classes by one class in each consecutive step, until AIC3 no longer decreases. The parameter estimates of the latent class model producing the lowest AIC3 are used to estimate π_{ii} (Equation 5.5). Figure 5.1 (left-hand panel) illustrates a series of latent class models having 1 to 4 latent classes. Probabilities π_j, π_{ij}, and test-score variance σ_X^2 are estimated by p_j, p_{ij}, and sample variance s_X^2, respectively.

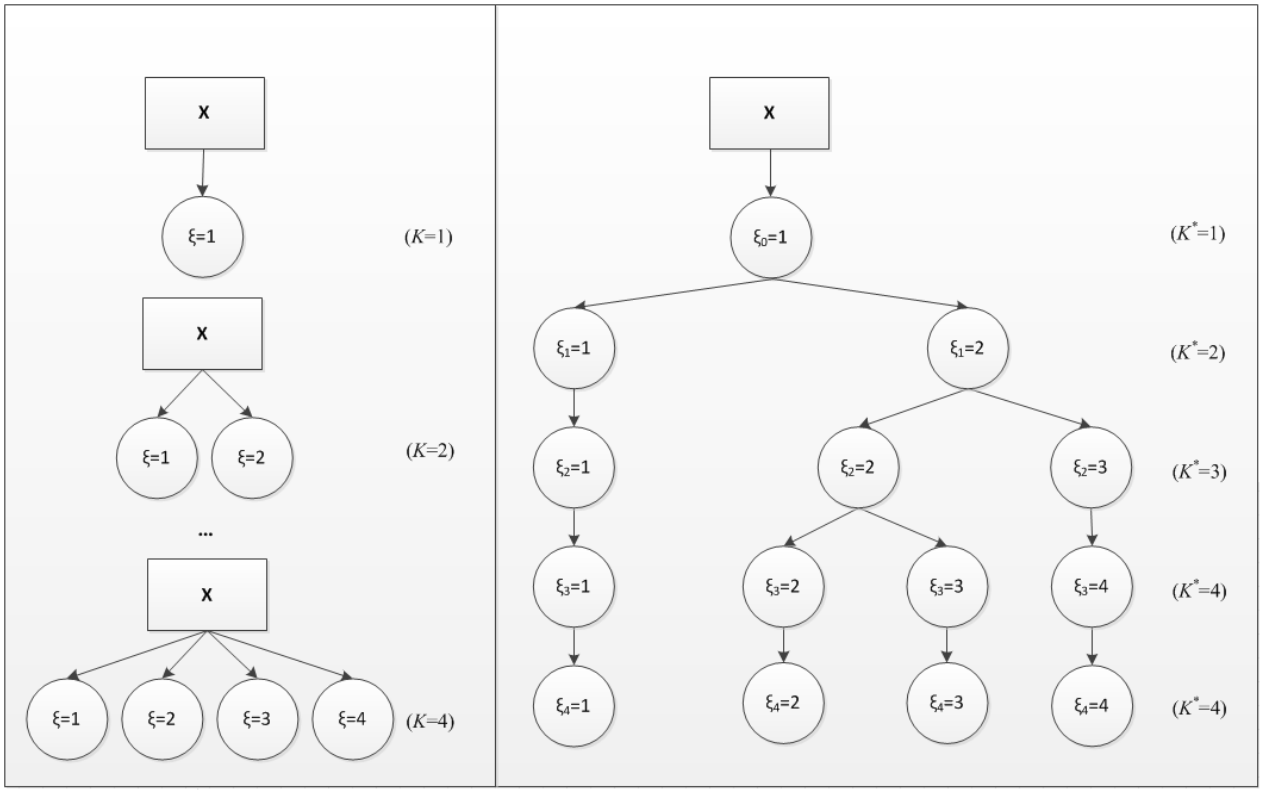

Figure 5.1. An illustration of the difference in the model-fit strategy for a traditional and a divisive latent class model. For the traditional latent class model in this example, four models were estimated to find the best fitting model. For the divisive latent class model, the number of latent classes is increased during the estimation procedure until sufficient model-fit is achieved.

The LCRC* Version of Method LCRC

The difference between LCRC* and LCRC resides in the estimation of π_{ii}. Method LCRC* uses a DLC model for estimating π_{ii}. The DLC model describes the joint probability

distribution $P(X_1 = x_1, \ldots, X_J = x_J)$ in terms of Equation 5.4, but requires K^* rather than K latent classes; typically, $K^* > K$. The big difference between the latent class model and the DLC model is the way the latent classes are obtained. In the latent class model all K latent classes are estimated simultaneously, whereas the DLC model involves a top-down clustering of respondents into latent classes using a series of one-class and two-class models. In the first step, a one-class and a two-class model are fitted to the entire sample. If the two-class model fits better than the one-class model, the sample is split into two latent classes; otherwise the sample is not split and the procedure stops. If the sample was split into two latent classes, in the subsequent steps, a one-class and a two-class model are fitted to the sample in each latent class. If the two-class model fits better than the one-class model, the latent class is further split into two latent classes; otherwise the latent class remains unaltered for the rest of the procedure. The procedure stops if none of the splits improves the fit relative to the local one-class models. Figure 5.1 (right-hand panel) illustrates the divisive top-down clustering of a sample into latent classes. Van der Palm et al. (2013b) provide details of the DLC model.

The criterion for splitting a latent class depends on the application at hand. For the computation of method LCRC*, we propose using AIC (Akaike, 1974); that is, only if the AIC value of the two-class model is less than the AIC-value of the one-class model, the latent class is split. AIC is a liberal relative fit index. Using AIC will produce relatively many splits (and, henceforth, a large K^*) compared to more conservative relative fit indices, such as AIC3. Because latent classes are not interpreted and only used as a tool to estimate densities as accurately as possible, a large number of latent classes is not problematic (also, see Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008).

Once the DLC model has been estimated, Equation 5.5 (with K^* classes rather than K) is used to estimate π_{ii}. As for LCRC, probabilities π_j, π_{ij}, and test-score variance σ_X^2 are estimated by p_j, p_{ij}, and s_X^2, respectively. Compared to method LCRC, method LCRC* is faster, requires less computer memory, requires only a single run, and is insensitive to human error because the researcher need not compare the fit of different latent class models.

Other Reliability Estimation Methods

We compared methods LCRC and LCRC* with the lower bounds coefficient alpha, coefficient lambda2, and GLB. The three lower bounds are related such that $\text{alpha} \leq \text{lambda2} \leq \text{GLB} \leq \rho_{XX'}$ (Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977). We

provide the equations for coefficients alpha and lambda2 and discuss the logic of GLB for which a single equation is not available but which results from an optimization procedure. Let σ_{jk} denote the covariance between items *j* and *k*. Coefficient alpha is defined as,

$$\text{alpha} = \frac{J}{J-1}\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sigma_X^2},$$

and coefficient lambda2 is defined as,

$$\text{lambda2} = \frac{\sum\sum_{j\neq k}\sigma_{jk} + \sqrt{\frac{J}{J-1}\sum\sum_{j\neq k}\sigma_{jk}^2}}{\sigma_X^2}.$$

GLB is obtained through a maximization process that finds the largest sum of the item-error variances given the data and the assumptions of classical test theory (e.g., Ten Berge & Sočan, 2004), thus creating the least favorable conditions for the reliability and, therefore, the process creates the lower bound of the interval $[GLB, 1]$ in which reliability $\rho_{XX'}$ is located. This is also the greatest value of the set of theoretical lower bounds, including alpha and lambda2. GLB tends to overestimate the reliability in small samples (Woodhouse & Jackson, 1977; Ten Berge & Sočan, 2004).

Method

We performed a simulation study to compare the bias and the accuracy of methods LCRC, LCRC*, alpha, lambda2, and GLB. We used the multidimensional two-parameter logistic model (M2PLM; Reckase, 1997) to generate 0/1 scores. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)$ denote the Q-dimensional latent variable vector; $\boldsymbol{\theta}$ has a Q-variate standard-normal distribution. Let ψ_{jq} denote the discrimination parameter of item j for latent variable q, and let δ_j denote the item location parameter. The M2PLM is defined as,

$$P(X_j = 1|\boldsymbol{\theta}) = \frac{\exp\left[\sum_{q=1}^{Q}\psi_{jq}(\theta_q - \delta_j)\right]}{1 + \exp\left[\sum_{q=1}^{Q}\psi_{jq}(\theta_q - \delta_j)\right]}. \quad (5.6)$$

The M2PLM and $\boldsymbol{\theta}$ were used to compute the population reliability, $\rho_{XX'}$ (Equation 5.1), and to generate the data. For the computation of $\rho_{XX'}$ by means of Equation 5.1, the population was defined by 10,000,000 randomly drawn $\boldsymbol{\theta}$s. First, for each $\boldsymbol{\theta}$ and each item, $P(X_j = 1|\boldsymbol{\theta})$ was computed using Equation 5.6. Second, for each $\boldsymbol{\theta}$, item scores were sampled from $P(X_j = 1|\boldsymbol{\theta})$ and test score X was computed as $X = \sum_{j=1}^{J} X_j$. The variance of X served as the

denominator on the right-hand side of Equation 5.1. Third, for each $\boldsymbol{\theta}$, true score T was computed as,

$$T|\boldsymbol{\theta}_n = \sum_{j=1}^{J} P(X_j = 1|\boldsymbol{\theta}_n).$$

The variance of T served as the numerator in the right-hand side of Equation 5.1.

The generation of each dataset started with randomly drawing for each simulee a latent-variable vector from a Q-variate standard normal distribution, yielding N vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. The correlations between the Q dimensions (denoted by r) were all equal. Subsequently, for each simulee Equation 5.6 was used to obtain the probability of a particular score for each of the J items, and the item scores were generated by random draws from a multivariate, uniform distribution.

Because the goal of the study was to compare reliability estimation in specific, relevant situations, we used a design with main effects instead of a full-factorial design. Table 5.1 shows the 14 conditions included in the simulation study. We extended the design Van der Ark et al. (2011) used and compared our results to theirs. Because Van der Ark et al. (2011) included a polytomous-data condition, we did this as well to allow a full comparison. For this condition, we used the multidimensional graded response model (De Ayala, 1994) to generate scores ranging from 0 to 4 (see Van der Ark et al., 2011, for the technical details). Van der Ark et al. (2011) defined a standard test condition as a reference point. This standard test condition was defined as: unidimensional data, equal discrimination parameters, 6 dichotomous items, and a sample size equal to 1000 (Table 5.1, top row). Subsequently, five conditions were investigated that each differed from the standard design with respect to one design factor (i.e., main effects: polytomous items, long test, small sample size, unequal discrimination parameters, and two-dimensional data). In addition to the six conditions Van der Ark et al. (2011) investigated, we investigated eight multidimensional-data conditions in which we varied the number of dimensions (2 or 3), the strength of the correlation between dimensions (0, .2 or .5), and the length of the test (6 or 18 items). The eight new conditions were designed to more closely approach an educational test setting; each item loads strongly on a primary dimension, and moderately on a secondary (and/or tertiary) dimension. The location parameters range from easy to difficult for each dimension. Methods LCRC, LCRC*, alpha, lambda2, and GLB were computed in each of the 14 design conditions, using 1000 replications in each condition.

Table 5.1: *Fourteen Design Conditions of the Simulation Study.*

Condition	Design Factors					
	Q	Q^*	r	Equal ψ_j	J	N
Design of Van der Ark et al. (2011)						
Standard	1	1	NA	yes	6	1000
Polytomous	1	1	NA	yes	6	1000
Long test	1	1	NA	yes	18	1000
Small N	1	1	NA	yes	6	200
Unequal ψ	1	1	NA	no	6	1000
2D-standard	2	1	.0	yes	6	1000
Additional design						
2D-0-short	2	2	.0	no	6	1000
2D-2-short	2	2	.2	no	6	1000
2D-5-short	2	2	.5	no	6	1000
2D-0-long	2	2	.0	no	18	1000
2D-2-long	2	2	.2	no	18	1000
2D-5-long	2	2	.5	no	18	1000
3D-5-short	3	3	.5	no	6	1000
3D-5-long	3	3	.5	no	18	1000

Note. Q = number of dimensions, Q^* = number of dimensions each items loads on, 2D = 2 dimensional data, 3D = 3 dimensional data, r = correlation between the dimensions (.0, unrelated, .2, weak, and .5, strong), J = number of items, N = sample size. As an example: 2D-5-short = condition with 2 dimensional data, a correlation of .5 between the dimensions, and six items.

Table 5.2 shows the parameter values of the M2PLM for the conditions referring to dichotomous items. For the conditions referring to 18 items, the parameter values of items 7 through 12 and 13 through 18 were identical to those of items 1 through 6. In the design used by Van der Ark et al. (2011), the two-dimension condition had a simple structure, whereas in the additional design all items loaded on all latent variables. The latent variables were either uncorrelated ($r = .0$), weakly correlated ($r = .2$) or strongly correlated ($r = .5$).

Table 5.2. *Item Parameters of the Multidimensional Two-Parameter Logistic Model.*

Item	Standard		Unequal ψ		2 Dimensions			2D			3D			
	ψ_j	δ_j	ψ_j	δ_j	ψ_{j1}	ψ_{j2}	δ_j	ψ_{j1}	ψ_{j2}	δ_j	ψ_{j1}	ψ_{j2}	ψ_{j3}	δ_j
1	1	-2.5	0.5	-2.5	1	0	-2.5	2	1	-2	2	1	1	-1.5
2	1	-1.5	2	-1.5	1	0	-1.5	2	1	0	2	1	1	1.5
3	1	-0.5	0.5	-0.5	1	0	-0.5	2	1	2	1	2	1	-1.5
4	1	0.5	2	0.5	0	1	0.5	1	2	-1.5	1	2	1	1.5
5	1	1.5	0.5	1.5	0	1	1.5	1	2	0	1	1	2	-1.5
6	1	2.5	2	2.5	0	1	2.5	1	2	1.5	1	1	2	1.5

The dependent variables were bias and accuracy. Let r_b denote a reliability estimate in replication b (b = 1, …, B). The bias was computed as

$$\text{bias} = \frac{1}{B}\sum_{b=1}^{B}(r_b - \rho_{XX'}).$$

To interpret the size of bias and accuracy values, we adopted the rules of thumb Van der Ark et al. (2011) used. Absolute bias was interpreted as follows: $|\text{bias}| < .001$ was considered negligible, $.001 \leq |\text{bias}| < .01$ small, $.01 \leq |\text{bias}| < .02$ medium, $.02 \leq |\text{bias}| < .05$ considerable, and $|\text{bias}| \geq .05$ large. To assess accuracy, the mean absolute error (MAE) was used, which is defined as

$$\text{MAE} = \frac{1}{B}\sum_{b=1}^{B}|r_b - \rho_{XX'}|.$$

MAE provides the error one can expect for a single dataset. MAE was interpreted as follows: $\text{MAE} < .002$ was considered negligible, $.002 \leq \text{MAE} < .02$ small, $.02 \leq \text{MAE} < .04$ medium, $.04 \leq \text{MAE} < .10$ considerable, and $\text{MAE} \geq .10$ large.

The study was conducted using R (R Core Development Team, 2012) and Latent GOLD (Vermunt & Magidson, 2008). All necessary syntax files are available from the first author. LCRC* was estimated using R and Latent GOLD (Vermunt & Magidson, 2008), LCRC, alpha, and lambda2 were estimated using the R package `mokken` (Van der Ark, 2007), and GLB was estimated using the R package `psych` (Revelle, 2013).

Results

First, the results for alpha, lambda2, and LCRC in the first six conditions (Table 5.3) were replications of Van der Ark et al. (2011). Our findings were very similar to theirs, which indicates that our simulations were carried out correctly.

Table 5.3. *Bias and Accuracy of Five Reliability Estimation Methods for 14 Conditions of the Simulation Study. Reliability, Bias and Accuracy Values (MAE) were Multiplied by 1000 to Improve Readability.*

		Bias				
Condition	$\rho_{XX'}$	LCRC	LCRC*	Alpha	Lambda2	GLB
Design of Van der Ark et al. (2011)						
Standard	464	-5	-7	-16	-7	28
Polytomous	765	-8	2	-14	-8	12
Long test	722	-2	1	-8	-3	45
Small *N*	464	-8	-7	-23	-7	66
Unequal ψ	424	-7	-6	-47	-32	13
2D-0-standard	315	3	-2	-80	-49	25
Additional design						
2D-0-short	640	-48	-26	-154	-97	-44
2D-2-short	680	-41	-15	-145	-88	-34
2D-5-short	724	-26	-1	-133	-79	-20
2D-0-long	855	-9	-7	-46	-27	16
2D-2-long	876	-10	-9	-43	-25	13
2D-5-long	897	-9	-9	-39	-23	10
3D-5-short	797	-14	-8	-88	-72	-9
3D-5-long	927	-11	-5	-34	-22	8
		Accuracy (MAE)				
Design of Van der Ark et al. (2011)						
Standard		23	22	24	21	33
Polytomous		11	9	15	11	14
Long test		10	10	12	10	45
Small *N*		46	46	50	47	73
Unequal ψ		28	27	48	36	25
2D-0-standard		32	34	80	51	37
Additional design						
2D-0-short		55	35	154	98	45
2D-2-short		45	23	145	88	34
2D-5-short		34	15	133	79	22
2D-0-long		10	8	46	27	16
2D-2-long		11	9	43	25	13
2D-5-long		9	9	39	23	10
3D-5-short		16	13	88	72	13
3D-5-long		11	6	34	22	8

For LCRC, the number of latent classes ranged between 2 and 7 with median equal to 3, and for LCRC*, the number of latent classes ranged between 1 and 8 with median equal to 4. Table 5.3 shows the true reliabilities, $\rho_{XX'}$, for the 14 simulation conditions, and the bias and the accuracy (MAE) of the five reliability estimation methods. LCRC* showed the smallest bias in all conditions. Only for condition 2D-0-short, bias was considerable and for 2D-2-short bias was medium. These two conditions (uncorrelated or weakly related dimensions and short tests) seem the least representative of educational tests.

Alpha and lambda2 showed either considerable or large negative bias for all conditions that included multidimensional data or data generated by IRT models with unequal discrimination parameters. For other conditions, lambda2 showed small negative bias and alpha showed small to considerable negative bias. The GLB showed positive bias (ranging from small to large) in some conditions and negative bias in others (ranging from small to considerable). Compared to the differences with respect to bias, the differences with respect to accuracy were relatively small across reliability estimation methods. For the design Van der Ark et al. (2011) used, the differences between the methods were small. For the additional design, LCRC and LCRC* were more accurate than alpha, lambda2, and the GLB.

Real-Data Examples

The five reliability estimation methods were used to estimate test-score reliability for available data from eight educational tests (see Table 5.4). The educational tests are bachelor-level examinations administered at the Tilburg School of Social and Behavioral Sciences, The Netherlands. Table 5.4 shows the courses, number of items (*J*), and sample size (*N*). For each test, we estimated test-score reliability using LCRC, LCRC*, alpha, lambda2, and GLB. The dimensionality of a dataset may affect bias and accuracy of reliability estimation. Hence, we investigated data dimensionality. Because item scores are categorical (correct-incorrect), factor analysis was not used (e.g., Kim & Mueller, 1978, p. 74). Instead, we used two other procedures to investigate data dimensionality. First, we used the automated item selection procedure in Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002), which partions a set of items into unidimensional scales, and is available in the R-package `mokken` (Van der Ark, 2007). Second, we used a scree-plot of the singular values of the principal axes in multiple correspondence analysis (Greenacre, 2007), which is available in the R package `ca` (Nenadic & Greenacre, 2007).

Table 5.4. Real Data Example; Application of the Five Reliability Estimation Methods to Real Educational Test Data, and Information on the Estimated Dimensionality of each Dataset.

Educational tests				Reliability estimation methods				Dimensionality	
Course	*J*	*N*	LCRC	LCRC*	Alpha	Lambda2	GLB	MSA	CA
Introduction to Statistics	20	617	.794	.801	.790	.795	.851	2	1
Experimental Methods	30	318	.769	.776	.765	.772	.864	2	1
Construction and Analysis of Questionnaires	29	306	.736	.743	.730	.740	.846	3	2
Introduction to Psychology	38	121	.563	.568	.546	.568	.747	3	4
Social Psychology	47	366	.677	.688	.677	.689	.838	4	4
Test Theory	23	248	.567	.577	.560	.581	.753	2	2
Introduction to Mathematics	23	54	.808	.821	.820	.835	.955	2	2
Causal Techniques	24	177	.768	.768	.771	.781	.883	2	2

Note. MSA = Mokken scale analysis: indicates the estimate number of scales (i.e., dimensions) present in each dataset.

Table 5.4 shows that Mokken scale analysis revealed multidimensional structures in each dataset. Correspondence analysis supported this finding in six datasets but suggested two

datasets were unidimensional. LCRC, LCRC*, alpha and lambda2 yielded small differences. LCRC* and lambda2 often produced higher estimates than LCRC and alpha. GLB produced much higher estimates than the other methods but it is likely that the results capitalized on chance. For the 'Introduction to Mathematics' and 'Causal Techniques' datasets, the difference between LCRC* and lambda2 was relatively large; lambda2 was .014 and .013 units higher than LCRC*, respectively. For the other courses, LCRC* and lambda2 were comparable.

Discussion

The most important finding of our simulation study is that LCRC* consistently yields the closest approximation to the true reliability. We recommend that practitioners use LCRC* to estimate test-score reliability whenever educational test data are multidimensional. For multidimensionality, bias differences between LCRC*, alpha and lambda2 were particularly pronounced.

Compared to LCRC, LCRC* was faster and easier to use. Our results were consistent with those found by Van der Ark et al. (2011) for the same test scenarios: Coefficients alpha and lambda2 are seriously biased when data are multidimensional and items have different discrimination parameters. The results of our simulation study reiterate that lambda2 produces a better reliability estimate than alpha and GLB. By definition, lambda2 is closer to the population reliability than alpha, and GLB is known to produce better estimates when the sample size exceeds 1000.

The real-datasets varied in size between 54 and 617, and estimation results certainly are subject to sampling error. Moreover, because the true reliability is unknown for real data we cannot know which reliability estimation method is closest to the true reliability. The real-data examples showed that lambda2 can be higher than LCRC* but the two datasets for which this result was found were small ($N = 54$ and $N = 177$), so sampling fluctuation may affect the estimated reliabilities. Given that lambda2 is a lower bound to the reliability and that LCRC* in general has little bias estimating test-score reliability, it may be reasonable to estimate both lambda2 and LCRC*, and report both.

Finally, we discuss a technical issue. For the estimation of the latent class models that are necessary to compute LCRC and LCRC*, we recommend trying many (approximately 100) different sets of starting values for the parameters (cf. Vermunt et al., 2008). This reduces the probability that the estimation algorithm ends up in a local maximum, which may have a small effect on the values of LCRC and LCRC*. In our R code (available upon request

from the first author), the number of different sets of starting values for the (divisive) latent-class model can be specified by the user.

Chapter 6

Conclusion and Discussion

The central theme of this thesis was the latent class model as a density estimation tool. A general finding of this thesis is that the performance of a latent class model as a density estimation tool is largely determined by how well a latent class model fits the data. More specifically, the model-fit strategy for the latent class model must be tailored to the specific application for which the density estimate is used. The aim of Chapter 2 was to investigate the performance the latent class model as an incomplete-data method (i.e., MILC). We compared the performance of MILC with methods maximum likelihood for incomplete data, multiple imputation using a log-linear model, and multivariate imputation using chained equations, and assessed the influence of sample size, number of variables, number of categories per variable, and complexity of associations on the bias and stability of the four incomplete-data methods. Based on the results of the simulation studies in Chapter 2, we conclude that MILC has further been established as a sound incomplete-data method. Furthermore, we conclude that it is essential to use a sufficiently large number of latent classes. Otherwise, the parameter estimates for the statistical model of interest may be severely biased. We do note that in practice one can never be exactly certain which number of latent classes is sufficient and, therefore, we recommend researchers and practitioners to be liberal when deciding on the number of latent classes an imputation model will contain. A rule of thumb, as suggested in Chapter 2, is to let the number of latent classes of the imputation model equal the sum of the number of categories of all variables included in the dataset. In the third study of Chapter 4 it was shown that letting the number of latent classes equal three times the total number of variables also resulted in small bias in the estimated parameters of the statistical model of interest. A future research topic is to investigate whether the good performance of MILC in combination with this rule of thumb generalizes to a wider range of incomplete-data scenarios.

The aim of Chapter 3 was to introduce an adaptation of the standard latent class model that trades model parsimoniousness for computational efficiency. We introduced the divisive latent class model as a density estimation method and showed that the computation time for a rather large example dataset was reduced by more than 87% in comparison to the computation

time for a standard latent class model. Furthermore, the generated data example showed that the divisive latent class model is capable of capturing complex associations. A crucial aspect of the divisive latent class model is the model-fit strategy or, more specifically, the decision rules for deciding whether each split sufficiently improves the model fit. Thus, the decision rules determine the final number of latent classes. An initial exploration was performed to assess the impact of different decision rules on the precision of the divisive latent class model as a density estimation method. We concluded that the divisive latent class model is able to yield a precise density estimate for a complex population model. However, further research is needed to assess the influence of factors such as sampling error, the number of variables, and complexity of associations. It is possible that more stringent decision rules are necessary to prevent the divisive latent class model from over fitting the data in the presence of sampling error and a larger number of variables.

We expect that the practical use of the divisive latent class model will increase in the future due to technological advancements such as the internet, which will stimulate the amount of available data to grow increasingly fast and individual datasets to become larger with respect to the number of variables and respondents, exacerbating the problem of excessive computation time for standard latent class models. Furthermore, the divisive latent class model is highly suited for parallel computing. Technological advancements in computer hardware during the last decade clearly show that multi-core processing is the future of computation. Thus, instead of increasing the computational speed of a single-core processor, hardware manufacturers are increasing the number of processing cores each processor contains. However, to make full and efficient use of multiple processing cores for estimating a statistical model, it is crucial that the processing load can be efficiently distributed across the different cores. In comparison to the traditional latent class model it is much easier and more efficient to make use of multi-core computation, or to even use distributed computation for the estimation of a divisive latent class model such as a cloud computing solution. Thus, estimating the divisive latent class model requires less time than the traditional latent class model because it is estimated sequentially, but also because it facilitates multi-core computation.

In Chapter 4 the divisive latent class model was investigated as an incomplete-data method. One remaining issue is that the performance of a divisive latent class model appears to be lower for dichotomous data than for polytomous data. Studies 2 and 3 of Chapter 4 clearly showed that the divisive latent class model can pick up complex associations, yet for dichotomous data, the divisive latent class model yielded relatively large bias in the estimate

of a three-variable interaction in the substantive model. Further research is required to investigate the cause of this difference in performance for different data types and, if possible, to adapt the method to resolve this issue. The nonparametric bootstrap must be used for the standard latent class model to obtain unbiased standard errors, however, in Chapter 4 we found that the divisive latent class model yielded sufficiently small bias in the standard errors without using the nonparametric bootstrap. This result was unexpected. Further research is required to investigate why the nonparametric bootstrap introduces bias in the results of the divisive latent class model.

In Chapter 5, the divisive latent class model was used to adapt the latent class based reliability estimation method introduced by Van der Ark et al. (2011), coined LCRC*. In the simulation study, it was found that LCRC* yielded a reliability estimate with the smallest bias and the highest accuracy for multidimensional educational test data. However, in the real-data examples the differences between LCRC* and the other reliability estimation methods were smaller. A future topic of research is to investigate what the cause is of these smaller differences. It is possible that the results were caused by the combination of a small sample size and a large number of items. Furthermore, in a pilot study, it was found that compared to LCRC* based on AIC3 using AIC improved the performance of LCRC* in terms of bias. It would be informative to investigate whether an even smaller penalty to the number of parameters would further increase the performance of LCRC*.

Finally, we discuss possible future research topics in a broader perspective. One limitation of the investigated latent class models for density estimation is that these concerned only categorical data. The usefulness of the presented applications could be further increased by developing generalizations allowing for the inclusion of continuous variables. Furthermore, in the social and behavioral sciences researchers frequently use multilevel or longitudinal datasets. It would be useful to investigate density estimation for such more complex data structures using more advanced models such as multilevel latent class models and latent Markov models. Lastly, it would be useful to investigate whether a newly developed divisive latent class modeling approach can be generalized for datasets containing both continuous and categorical variables, and for datasets with a multilevel or longitudinal structure.

References

Agresti. A. (1990). *Categorical data analysis.* New York: Wiley.

Akaike, H., (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. *SAS Users Group International Conference*, pp. 10-13. Philadelphia, PA.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides, & S. E. Schumacker (Eds.), *Advanced Structural Equation Modeling,* pp. 243-277. Mahway, NJ: Erlbaum.

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine, 26,* 1368-1382.

Bentler, P. A., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika, 45,* 249-267.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. M. (with collaboration of Light, R. J., & Mosteller, F.). (1975). *Discrete multivariate analysis: theory and practice.* Cambridge, MA: MIT Press.

Blenkner, M., Bloom, M., & Weber, R. (1974). *Final report: Protective services for older people*. Ohio: Benjamin Rose Institute.

Bouguila, N., & ElGuebaly, W. (2009). Discrete data clustering using finite mixture models. *Pattern Recognition, 42,* 33-42.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-fisher information matrix. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and classification, concepts, methods and applications*, pp. 40-54. Berlin: Springer.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences.* Hoboken, NJ: Wiley.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 10*, 255-282.

Cronbach, L. J., Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified parallel tests. *Educational and Psychological Measurement, 25,* 291-312.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18,* 155-170.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B, 39*, 1-38.

Drechsler, J., & Rassler, S. (2008). Does convergence really matter? In C. H. Shalabh (Ed.), *Recent advances in linear models and related areas*, pp. 341-355. *Essays in honour of Helge Toutenburg*. Berlin: Springer.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* London: Chapman & Hall.

European Commission (2006). *EU research on social sciences and humanities. Psychological contracts across employment situations: PSYCONES.* Retrieved May 16, 2013, from http://cordis.europa.eu/documents/documentlibrary/100123961EN6.pdf.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis.* London: Arnold

Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., Rubin, D. B., & Schafer, J. L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS Health Examination Surveys. *Proceedings of the Annual Research Conference*, pp. 257-266. Washington: Bureau of the Census.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of American Statistical Association, 77*, 270-278.

Gebregziabher, M., & DeSantis, S. M. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference, 140,* 3252-3262.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika, 60,* 179-192.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215-231.

Graham, J, W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research*, pp. 1-29. Thousand Oaks, CA: Sage.

Greenacre, M. (2007). *Correspondence analysis in practice.* London: Chapman & Hall/CRC.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Hagenaars, J. A. P. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis.* Newbury Park, CA: Sage.

Hagenaars, J. A. P., & McCutcheon, A. L. (Eds.) (2002). *Applied latent class analysis.* Cambridge, UK: Cambridge University Press.

Hoijtink, H., & Notenboom, A. (2004). Model based clustering of large datasets: Tracing the development of spelling ability. *Psychometrika, 69,* 481-498.

Horton, N. J., Lipsitz, S. P., & Parzen, M. A. (2003). Potential for bias when rounding in multiple imputation. *American Statistician*, *57*, 229-232.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., & Herring, A. H. (2005). Missing data methods in generalized linear models: a comparative review. *Journal of the American Statistical Association, 100*, 332-346.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika, 42,* 567-578.

Kamata, A., Turhan, A., & Darandari, E. (2003). Multidimensional composite scale scores. Paper presented at the *Annual Meeting of the American Educational Research Association*, Chicago, IL. Retrieved October 17, 2012, from http://mailer.fsu.edu/~akamata/papers/MD_rel_paper.pdf.

Kang, J. D. Y, & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22,* 523-539.

Kaufman L., & Rousseeuw, P.J. (1990). *Finding groups in data*. New York, Wiley.

Kim, J.- O. & Curry J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research, 6*, 206-240.

Kim, J.- O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. Newbury Park, CA: Sage.

Keel, P., Fichter, M., Quadflieg, N., Bulik, C., Baxter, M., Thornton, L. et al. (2004). Application of latent class analysis to empirically defined eating disorder phenotypes. *Archives of General Psychiatry, 61,* 192-200.

Klebanoff, M. A., & Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. A*merican Journal Epidemiology, 168,* 355-357.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement, 21,* 337-348.

Kurian, A., Gallagher, S., Cheeyandira, A., & Josloff, R. (2010). Predictors of in-hospital length of stay after laparoscopic ventral hernia repair: results of multivariate logistic regression analysis. *Surgical Endoscopy, 24,* 2789-2792.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction*, pp. 361-412. Princeton: Princeton University Press.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1,* 98-107.

Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis, 19*, 173-187.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data (2nd ed.)*, pp. 266-291. New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luciano, J. V., Ayuso-Mateos, J. L., Aguado J., Fernandez, A., Serrano-Blanco, A., Roca M, & Haro, J. M. (2010). The 12-item World Health Organization disability assessment schedule II (WHODAS II): a nonparametric item response analysis. *BMC Medical Research Methodology*, *10* (45), 1-9.

Magidson, J., & Vermunt, J. K., (2004). Latent class models. In D. Kaplan (Eds.), *Handbook of quantitative methodology for the social sciences*, pp. 175-198. Newbury Park, NJ: Sage.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, NJ: Sage.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. New York/Berlin: De Gruyter.

Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden*, *9* (28), 115-126.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873-900.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus 6.1*. Los Angeles: Muthén & Muthén.

Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional Graphics: The ca Package. *Journal of Statistical Software, 20* (3), 1-13.

Neyman, J., & Pearson, E. S. (1933*).* On the problem of most efficient tests of statistical hypotheses*. Philosophical Transactions of the Royal Society of London Series A*, 231: 289-337.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343-355.

R Development Core Team (2012). R: A language and environment for statistical computing [computer programming language]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health, 25*, 99-117.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375-385.

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25,* 69-76.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287-297.

Revelle, W. (2013). psych: Procedures for Personality and Psychological Research (Version 1.3.2). Northwestern University: Evanston. Retrieved April 15, 2013, from http://CRAN.R-project.org/package=psych.

Richards, G. (2010). The traditional quantitative approach. Surveying cultural tourists: Lessons from the ATLAS Cultural Tourism Research Project. In G. Richards, & W. Munsters (Eds.), *Cultural Tourism Research Methods*, pp. 13-32. Wallingford, UK: CABI

Rindskopf, D., & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine, 5,* 21-27.

Royston, P. (2009). Multiple imputation of missing values: Further update of ICE, with an emphasis on categorical variables. *The Stata Journal, 9,* 466-477.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

SAS Inc. (2011). *SAS for Windows*. Cary, NC: SAS.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J.A., & Rubin, D. B. (1996). The NHANES III multiple imputation project. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, retrieved April 24, 2012, from http://www.amstat.org/sections/srms/Proceedings/papers/1996_004.pdf.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52,* 79-97.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage

SPSS Inc. (2011). *SPSS 19 for Windows.* Somers, New York: IBM.

StataCorp. (2011). *Stata statistical software: Release 12.* College Station, TX: StataCorp.

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika, 46,* 201-213.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613-625.

Ueda, N., & Nakano, R. (2000). EM algorithm with split and merge operations for mixture models. *Systems and Computers, 31*, 930-940.

Vansteelandt, S., Carpenter, J., & Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology, 6,* 37-48.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16,* 219-242.

Van Buuren, S., Brand, P. L., Groothuis-Oudshoorn, K., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76,* 1049-1064.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45* (3), 1-67.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20* (11), 1-19.

Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent-class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380-392.

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2013a). A comparison of incomplete-data methods for categorical data. *Statistical Methods in Medical Research*, in press.

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2013b). Divisive latent class modeling as a density estimation method for categorical data. Manuscript submitted for publication.

Van Hattum, P., & Hoijtink, H. (2009). Market segmentation using brand strategy research: Bayesian inference with respect to mixtures of log-linear models. *Journal of Classification, 26,* 297-328.

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data*. Tilburg: Department of Methodology and Statistics, Tilburg University, The Netherlands.

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. S. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The Sage Encyclopedia of Social Science Research Methods*, pp. 549-553. Thousand Oaks, CA: Sage.

Vermunt, J. K., & Magidson, J. (2008). *LG-syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 38,* 369-397.

Wang, H. X., Luo, B., Zhang, Q. B., & Wei, S. (2004). Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm. *Pattern Recognition Letters, 25,* 1799-1809.

White, I. R., Royston, P., & Wood, A. M. (2010). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30,* 377-399.

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42,* 579-591.

Zhu, B., Walter, S. D., Rosenbaum, P. L., Russell, D. J., & Raina, P. (2006). Structural equation and log-linear modeling: a comparison of methods in the analysis of a study on caregivers' health. *BMC Medical Research Methodology, 6* (49), 1-14.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33-49.

Summary

The topic of this thesis is the investigation and application of the latent class model as a *density estimation tool*. Typically, the latent class model is used for the identification of meaningful subgroups in the data. More recently, the model has also been used to estimate the multivariate distribution of a set of categorical variables based on sample data. When used as a density estimation tool, the parameters of the latent class model should contain as much information as possible about the associations among the categorical variables in a dataset. The latent class model estimates complex multivariate densities by means of a mixture of simple univariate multinomial densities. Within each latent class, it is assumed that the variables are statistically independent, which is also known as the local independence assumption. Because of local independence, the latent class model has a relatively simple model structure. If the latent class model is used as a tool rather than a model for understanding latent group structure, parameters need not—and usually cannot—be interpreted, the number of latent classes typically is large, and the latent class model need not be identifiable. The only criterion that matters for density estimation is whether the latent class model sufficiently captures all relevant associations.

Density estimation is useful, for example, when researchers and practitioners wish to smooth large sparse contingency tables, estimate incomplete data using multiple imputation, and estimate test-score reliability. As a density estimation tool, the latent class model is particularly useful because the model can handle a large number of variables. The log-linear model can also be used for density estimation, but its applicability is limited due to computational problems: Unless a very simple association structure is specified, a log-linear model can only be estimated for small numbers of variables. For density estimation, it is common practice to use a saturated log-linear model, because it captures all associations that are theoretically possible. However, for a saturated log-linear model computational problems emerge even sooner than for other models and this inspired the development of the latent class model as a density estimation tool.

This thesis addressed the following four research questions. First, previous research investigated the performance of the latent class model as an incomplete-data method. However, the latent class approach had not yet been compared to multivariate imputation

using chained equations, and the influence of sampling error, sample size, number of variables, number of response categories, and complexity of associations on the performance of the latent class model as an incomplete-data method had not yet been investigated. Second, the model-fit strategy for a standard latent class model may require excessive computation time and the question arose whether a more efficient algorithm could be developed that reduces computation time. Third, the more efficient estimation algorithm for latent class models had not yet been applied to the problem of incomplete data. And lastly, previous research showed that the latent class model can be used for test-score reliability estimation, yielding small bias relative to the population reliability. However, it was unknown whether the bias result would generalize to a wider range of scenarios involving multidimensional data. In addition, it was unknown how the more efficient estimation algorithm for the latent class model would perform as an adaptation of the standard latent class approach. In Chapters 2, 3, 4, and 5 we address these questions.

In Chapter 2, we studied four methods for handling incomplete categorical data in statistical modeling: (1) maximum likelihood estimation of the statistical model with incomplete data, (2) multiple imputation using a log-linear model, (3) multiple imputation using a latent class model, (4) and multiple imputation by means of chained equations. Each method has advantages and disadvantages, and it was unknown which method should be recommended to practitioners. We reviewed the merits of each method and investigated their effect on the bias and stability of parameter estimates and on the bias of the standard errors. We found that multiple imputation using a latent class model with many latent classes was the most promising method for handling incomplete categorical data, especially when the number of variables used in the imputation model is large.

In Chapter 3, we introduced a divisive latent class (DLC) model as a density estimation tool that may offer several advantages in comparison to a standard latent class model. When using a latent class model for density estimation, a considerable number of increasingly larger latent class models may have to be estimated before sufficient model-fit is achieved. A DLC model consists of a sequence of small latent class models. Therefore, a DLC model can be estimated much faster than a standard latent class model and can easily utilize multiple processor cores, meaning that the model is widely applicable and practically convenient. We described the algorithm of fitting a DLC model, and discussed the various settings that indirectly influence the precision of a DLC model as a density estimation tool. These settings were illustrated using a synthetic data example, and the best performing algorithm was applied to a real-data example. The generated data example showed that, using

specific decision rules, a DLC model correctly models complex association among categorical variables.

In Chapter 4, we investigated the performance of the DLC model as an incomplete-data method. Relatively few incomplete-data methods are available for categorical data and the methods that are available suffer from serious practical problems. Maximum likelihood estimation for incomplete data and multiple imputation using a log-linear model are the two most frequently used methods for incomplete categorical data. Yet, due to computational problems maximum likelihood estimation for incomplete data and multiple imputation using a log-linear model can handle only a few variables. Previous research showed that multiple imputation using a latent class model has a performance in terms of bias and stability of parameter estimates comparable to that of maximum likelihood estimation for incomplete data and multiple imputation using a log-linear model. However, the required model-fit strategy for multiple imputation using a latent class model may pose an obstacle to its practical usefulness. Multiple imputation using a DLC model solves the problems of maximum likelihood estimation for incomplete data, multiple imputation using a log-linear model, and multiple imputation using a latent class model: The DLC method can handle a very large number of variables, is easier to use, and much faster to compute. However, the statistical properties of multiple imputation using a DLC model are unknown. We used three studies to compare the performance of the DLC model as incomplete-data method with several commonly used incomplete-data methods. Results showed that the DLC model generally has a performance comparable to that of the standard latent class approach. However, the performance of a divisive latent class model appears to be lower for dichotomous data than for polytomous data. Further research is required to investigate the cause of this difference in performance for different data types and, if possible, to adapt the method to resolve this issue.

In Chapter 5, we investigated two latent class approaches to reliability estimation for multidimensional educational test data. Most items in an educational test require for their solution multiple abilities, skills, and knowledge of several topics and, therefore, when administered to students yield multidimensional data. Reliability estimation methods for multidimensional data are available but suffer from several practical problems. We proposed the adapted latent class reliability coefficient that solves these problems and is particularly suited for multidimensional data. Results showed that the adapted latent class reliability coefficient produces a less biased reliability estimate than other methods in a wide range of scenarios involving multidimensional data.

Samenvatting (Summary in Dutch)

Het onderwerp van dit proefschrift is het latente klassenmodel als schatter van kansverdelingen en toepassingen daarvan. Normaliter wordt het latente klassenmodel gebruikt voor de identificatie van betekenisvolle subgroepen op basis van categorische data. Echter, meer recent is het model ook gebruikt om de multivariate verdeling van een set categorische variabelen te schatten op basis van een steekproef uit die populatie. Bij dit type toepassing zouden de parameters van het latente klassenmodel zoveel mogelijk informatie moeten bevatten omtrent de relaties tussen de categorische variabelen in een dataset. Een latente klassenmodel schat complexe kansverdelingen door middel van een gewogen gemiddelde van relatief simpele univariate kansverdelingen, in de Engelstalige literatuur is dit type model ook wel bekend als een 'mixture model' en heeft de aanname dat de antwoorden die respondenten binnen één klasse geven statistisch onafhankelijk zijn. Vanwege deze lokale onafhankelijkheid heeft het latente klassenmodel een relatief simpele structuur. Als het latente klassenmodel gebruikt wordt als schatter van kansverdelingen hoeven de parameters niet inhoudelijk geïnterpreteerd te worden en dit is vaak ook niet mogelijk. Verder is het gespecificeerde aantal latente klassen meestal groot en hoeft het model niet geïdentificeerd te zijn. Het enige belangrijke criterium bij schatting van kansverdelingen door middel van een latente klassenmodel is de mate waarin alle relevante relaties tussen de variabelen correct beschreven worden door het model.

Het schatten van kansverdelingen is nuttig voor vele doeleinden zoals het verbeteren van de analyse van grote kruistabellen met veel nulfrequenties, het schatten van incomplete data met behulp van multipele imputatie en het schatten van de betrouwbaarheid van testscores. Het latente klasse model is bijzonder geschikt voor het schatten van kansverdelingen omdat het model geschat kan worden voor datasets met daarin een groot aantal variabelen. Het log-lineaire model kan ook gebruikt worden om kansverdelingen te schatten, maar toepassing van dit model is beperkt vanwege computationele problemen: Als men niet een relatief simpele associatiestructuur specificeert kan het log-lineaire model alleen geschat worden voor datasets met daar in een klein aantal variabelen. Het is bovendien ook gebruikelijk om een verzadigd log-linear model te gebruiken voor het schatten van kansverdelingen omdat het model dan alle theoretisch mogelijke relaties correct kan beschrijven. In het geval van een verzadigd log-lineair model is er nog sneller sprake van

computationele problemen en dit inspireerde de ontwikkeling van het latente klassenmodel voor schatting van kansverdelingen.

In dit proefschrift behandelen we vier onderzoeksvragen. 1. In eerder onderzoek is het latente klassenmodel onderzocht als methode voor het hanteren van incomplete data. Echter, de latente klassen methode had men nog niet vergeleken met 'multiple imputation using chained equations' en het was nog niet onderzocht wat de invloed is van steekproeffluctuatie, steekproefgrootte, aantal variabelen, aantal antwoordcategorieën en complexiteit van relaties op de werking van het latente klassenmodel als methode om incomplete data te hanteren. 2. De gebruikelijke procedure om een goed passend latente klassenmodel te vinden kan erg veel rekentijd vereisen en daardoor ontstond de vraag of een schattingsalgoritme met een hogere efficiëntie ontwikkeld kon worden, dat deze rekentijd vermindert. 3. Dit schattingsalgoritme met een hogere efficiëntie was nog niet toegepast op het probleem van incomplete data. 4. Uit eerder onderzoek is gebleken dat het latente klassenmodel gebruikt kan worden om de betrouwbaarheid van testscores te schatten met een kleine gemiddelde afwijking ten aanzien van de populatiebetrouwbaarheid. Het was echter nog niet bekend of dit resultaat zou generaliseren naar een groter bereik van scenario's die multidimensionele data betreffen. Daarnaast was het ook niet bekend of het schattingsalgoritme met een hogere efficiëntie goed zou werken als aanpassing van de standaard latente klassen methode voor betrouwbaarheidsschatting. In hoofdstukken 2, 3, 4 en 5 behandelen we deze vragen.

In hoofdstuk 2 bestudeerden we vier methoden voor het statistisch modelleren van incomplete categorische data: (1) 'maximum likelihood estimation of the statistical model with incomplete data', (2) 'multiple imputation using a log-linear model', (3) 'multiple imputation using a latent class model', (4) en 'multiple imputation by means of chained equations'. Alle vier de methoden hebben voor- en nadelen en het was nog niet bekend welke methode aangeraden zou moeten worden voor gebruik in de praktijk. We onderzochten wat het effect was van elke methode op de 'bias' en stabiliteit van parameter schattingen en op de 'bias' van de standaardfouten. Uit de resultaten bleek dat een latente klassenmodel met veel latente klassen de meest veelbelovende methode was voor het hanteren van incomplete categorische data, met name voor datasets met een groot aantal variabelen.

In Hoofdstuk 3 introduceerden we het 'divisive latent class (DLC)' model dat als schatter van kansverdelingen mogelijk meerdere voordelen biedt in vergelijking met het standaard latente klassenmodel. Als het latente klassenmodel gebruikt wordt voor het schatten van kansverdelingen en het best passende model moet worden gevonden, dan moet men vaak een groot aantal latente klassenmodellen schatten waarvan elk opvolgend model steeds meer

latente klassen bevat. Een DLC model bestaat uit een serie van latente klassenmodellen, elk met slechts een klein aantal klassen. Hierdoor kan een DLC model sneller geschat worden dan een standaard latente klassenmodel en kan men gemakkelijker meerdere kernen van een computer processor aanspreken. Vanwege deze eigenschappen is het DLC model breder toepasbaar en gebruikersvriendelijker. We beschreven het schattingsalgoritme van het DLC model evenals de verschillende instellingen die indirect de accuratesse van een DLC model als schatter van kansverdelingen beïnvloedt. De verschillende instellingen van het DLC model werden geïllustreerd aan de hand van een voorbeeld met gegenereerde data en het meest accurate algoritme werd toegepast op echte data. Het voorbeeld met gegenereerde data liet zien dat op basis van specifieke beslisregels een DLC model in staat is om complexe relaties tussen categorische variabelen correct te beschrijven.

In Hoofdstuk 4 onderzochten we het DLC model als methode voor het hanteren van incomplete data. Relatief weinig methodes zijn beschikbaar voor het hanteren van incomplete categorische data en het gebruik van de beschikbare methoden gaat gepaard met ernstige praktische problemen. 'Maximum likelihood estimation for incomplete data' en 'multiple imputation using a log-linear model' zijn de twee meest gebruikte methoden voor het hanteren van incomplete categorische data. Echter, vanwege computationele problemen kunnen deze twee methoden slechts gebruikt worden voor datasets die een klein aantal variabelen bevatten. Voorgaand onderzoek liet zien dat 'multiple imputation using a latent class model' even goed presteert als 'maximum likelihood estimation for incomplete data' en 'multiple imputation using a log-linear model'. Echter, de benodigde procedure om het best passende latente klassenmodel te vinden kan wederom een drempel zijn. 'Multiple imputation using a DLC model' lost de problemen van 'maximum likelihood estimation for incomplete data', 'multiple imputation using a log-linear model' en 'multiple imputation using a latent class model' op: De DLC methode kan datasets met een groot aantal variabelen hanteren, is gemakkelijker in gebruik en sneller om te berekenen. Echter, de statistische eigenschappen van 'multiple imputation using a DLC model' waren nog niet onderzocht. Door middel van drie studies hebben we het functioneren van het DLC model als methode om incomplete data te hanteren onderzocht. Uit de resultaten kwam naar voren dat het DLC model over het algemeen even goed functioneert als het standaard latente klassenmodel. Het moet wel opgemerkt worden dat het DLC model minder goed lijkt te functioneren voor dichotome variabelen dan voor polytome variabelen. Verder onderzoek is nodig om the achterhalen wat de oorzaak is van dit verschil in het functioneren van het DLC voor deze twee type data en, indien mogelijk, om de methode aan te passen zodat dit probleem verholpen wordt.

In Hoofdstuk 5 onderzochten we twee latente klassen methodes voor het schatten van testscorebetrouwbaarheid voor multidimensionele testdata verkregen met onderwijskundige toetsen. Voor de meeste items in een onderwijskundige toets geldt dat er meerdere vaardigheden en kennis van verscheidene onderwerpen nodig is om tot een correct antwoord te komen. Als een dergelijke toets wordt voorgelegd aan studenten dan verkrijgt men om die reden dan ook multidimensionele data. Er zijn wel methoden beschikbaar om de betrouwbaarheid te schatten voor multidimensionele data, maar het gebruik van deze methoden wordt beperkt door praktische problemen. Wij introduceerden een betrouwbaarheidscoëfficiënt gebaseerd op het aangepaste latente klassenmodel. Uit de resultaten komt naar voren dat deze aangepaste coëfficiënt gemiddeld minder afwijkt van de populatiebetrouwbaarheid dan andere methoden, voor een groot aantal scenario's die multidimensionele data betreffen.

Dankwoord

Na vier intensieve, leerzame en mooie jaren is een nieuwe mijlpaal dan bereikt: Het schrijven van mijn proefschrift is afgerond. Allereerst wil ik mijn begeleiders Jeroen Vermunt, Klaas Sijtsma en Andries van der Ark bedanken. Jeroen bedank ik graag voor zijn expertise en passie omtrent programmeren en categorische data analyse. Jeroen ontdekte al snel dat ik ook erg enthousiast werd van programmeren en het was vervolgens gelukkig ook mogelijk om extra tijd en aandacht te geven aan dit aspect van het project. Klaas en Andries bedank ik graag voor de prettige samenwerking en het ondersteunen en bewaken van het schrijfproces. Dankzij hun expertise heb ik veel bij mogen leren op het gebied van de psychometrie en vele gerelateerde onderwerpen.

Ik dank Greg Richards voor het beschikbaar stellen van de dataset die gebruikt is in hoofdstuk 3, Jeroen de Jong voor de dataset die gebruikt werd in Hoofdstuk 4 en Luc van Baest, Wilco Emons, Guy Moors, Marcel van Assen en Ilja van Beest voor het beschikbaar stellen van de tentamen-datasets die gebruikt werden in hoofdstuk 5. Ook dank ik de leden van de extended-VICI groep voor het becommentariëren van hoofdstuk 4: Fetene, Verena, Erwin, Dereje, Geert, Margot, Zsuzsa en Daniel. Verder ben ik al mijn collega's bij het MTO departement heel erg dankbaar voor het prettige samenwerken en natuurlijk ook alle gezellige gesprekken. In het bijzonder bedank ik Milosh Kankarash, Marcel van Assen, Daniel Oberski, Ruslan Jabrayilov, Renske Kuijpers, Erwin Nagelkerke, Hendrik Straat, Ruud van Keulen, Peter Kruyen, Pieter Oosterwijk en Joost van Ginkel. Ook ben ik erg blij met de vele prachtige momenten die ik op de verschillende congressen heb mogen beleven, vaak samen met IOPS collega's. Daarbij denk ik aan de karaoke avond in Hong Kong (met Renske, Dylan en Joost), de extreme hitte en voetbal wedstrijden in Georgia en de prachtige cultuur en omgeving in Santiago de Compostella. Ik wil ook Gerko Vink bedanken voor alle interessante gesprekken over o.a. missende data en MICE. Verder heb ik met veel plezier deelgenomen aan de MTO band waarmee we allerlei covers te gehoren brachten en zullen gaan brengen: Paulette, Michèle, Wilco, Joris en Maurits, you guys rock! Ook wil ik Mandy, Marco en Anja bedanken voor hun steun gedurende vele mooie jaren.

Een groot deel van mijn promotie-traject beleefde ik samen met Zsuzsa en Margot, mijn kamergenoten. Jullie begrijpen als geen ander hoe het is om te worstelen met problemen

omtrent latente klassen analyse die soms toch wel erg manifest werden. Naast alle steun zijn jullie ook nog eens bereid om als paranimfen op te treden, waar ik erg blij mee ben! Ik dank ook graag familie Ippel, mijn schoonfamilie, voor hun support.

Ook bedank ik graag mijn familie, zonder wie al deze jaren van studie en werk niet mogelijk waren geweest: Mijn vader Willem van der Palm, mijn moeder Judith de Bruijn en mijn broers en zussen Marjolein, Sylvia, Renata, Christiaan en David. In het bijzonder ben ik mijn vader dankbaar wiens kalmte, vastberadenheid en wilskracht altijd tot groot voorbeeld zijn geweest voor mij en van grote betekenis blijven in alles wat ik onderneem.

Tot slot richt ik me tot Lianne Ippel: Bedankt voor al je positieve energie en aanmoedigingen. Wat fijn om elkaar zo goed te kunnen begrijpen, wat MTO betreft, maar ook wat betreft al het moois buiten die wereld.